Fundamentals of Psychological Measurement

Reagan Brown

Psychological Measurement: Why We Care

Measurement is important to all fields of science.

Without accurate measurement psychology would not be a science.

The stakes have never been higher.

The Importance of Measurement

Despite what is said by most physicists and my wife (with her fancy chemistry minor), psychology is a science. But what does it mean to be a science? It means, among other things, proposing testable hypotheses and then actually testing them. The key words are *testable* and *testing*. Even if we argued all day about the merits of an untestable hypothesis, we will never be any closer to learning whether this hypothesis conforms with reality. Moreover, it's not enough to perform some kind of a test of the hypothesis – we must perform a good test of it.

You may already be familiar with the basic elements of research design (e.g., random assignment to groups, experimental control). The design of the study is one important component of an adequate test of our hypothesis. Measurement is the other. Consider the following quotation.

> The government are very keen on amassing statistics – they collect them, raise them to the n^{th} power, take the cube root, and prepare wonderful diagrams. But what you must never forget is that every one of these figures comes in the first instance from the village watchman, who just puts down what he [expletive deleted] pleases. (Stamp, 1929, p. 258)

Thus, we must measure our variables with care. For what value is there in analyzing data that is no better than the village watchman's? Just as a flawed research design will prohibit us from performing a meaningful test of our hypothesis, so will flawed measurement. Without quality measurement we can never test our hypotheses, and psychology ceases to be a science. The issues are that simple. And they are that important.

Testing Terminology

It is unfortunate, but any discussion of psychological measurement requires an understanding of a few key concepts. These are important, and we will be using them for the rest of the book.

Item: Any test stimulus that produces a single response. Each response is an observable behavior. For example, the GRE Verbal section has a paragraph followed by five questions about that paragraph. Each of these five questions are individual items.

However, an item may not explicitly ask a question. Many tests of attitudes or personality make statements followed by a scale indicating how much one agrees with the statement. A common item type is as follows:

I am absolutely thrilled to learn about psychological measurement.

A. I strongly agree

B. I agree

C. I neither agree nor disagree

D. I disagree

E. I strongly agree

Test: A collection of items. Although some tests have only one item, most tests are composed of multiple items. The items can take a variety of forms (e.g., a paper and pencil test, an interview, an observation of children on a playground, performance on a flight simulator, ratings of someone's job performance, weight in pounds on a scale).

Also integral to a test is a system for administering the test and assigning points to the responses. The test must be administered the same way to each test taker. If the test has a 30 minute

time limit, then every test taker must finish the test in the allotted time. The same responses to a test item should be scored in the same way. If "C" is the correct answer to Item 27, then everyone who answers "C" should receive a point.

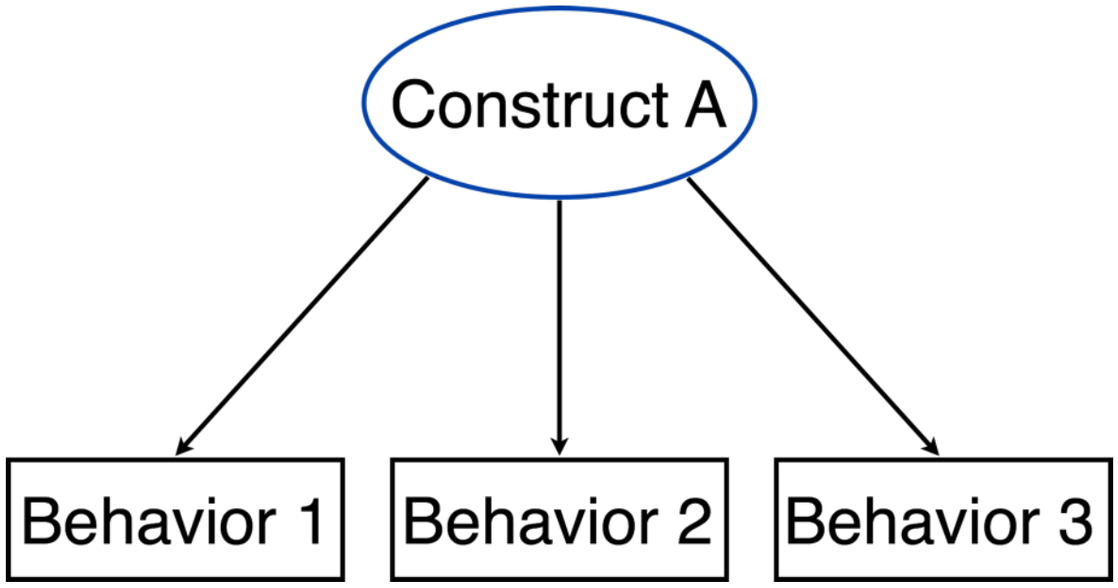
Finally, tests measure a sample of behaviors. We likely cannot measure every behavior relevant to the purpose of the test. Not every math operation can be measured on a math test. Performance on the sample of behaviors measured by the test can be interpreted as being representative of this larger domain of relevant behaviors. Or it can be interpreted as representing something beyond a set of observable behaviors (more on this below).

Construct: A hypothetical, unobservable cause for an observable behavior. Consider the construct of intelligence. We can't see or directly measure someone's intelligence (thus, the unobservable part). We can measure head size, brain mass, or performance on a math test, but none of these are direct measures of intelligence. Because we can't see intelligence, it is hypothetical. We think it exists, but we don't know for sure. Because we can't measure the construct directly, we measure the presumed effects of the construct by measuring observable behaviors. Why can a person correctly answer all of the items on the intelligence test? It must be because she is smart. Her intelligence is causing her to behave in this manner. By assuming that constructs cause behavior, we can use performance on the items that appear on the intelligence test to infer the test taker's actual level of intelligence.

So that's the primary definition of construct. And to some researchers (e.g., Ebel, 1975), it's the only definition. But there is an additional and equally valid definition. That is, instead of a construct causing a set of behaviors, the other view is that a construct *is* a set of behaviors (Guion, 1977). From this viewpoint a construct is not something unobservable; it's just a collection of re-

lated behaviors (not just any set of behaviors, but a meaningful set – as in they are all a part of the same activity). Driving a car can be defined as a set of behaviors (e.g., parallel parking, left turns, etc.) with no reference to underlying causes or abilities. On a driving test, we are not trying to determine whether a person has a high ability at some unobservable driving construct, we just want to determine if the person can successfully perform the same set of behaviors that actual drivers perform. In this case, we generalize from performance on the sample of behaviors measured on the test to the larger domain of relevant behaviors. To use the driver's test example again, we can't include every driving behavior on our road test – the test would too long. So we make our test a sample of behaviors from this larger domain of all possible driving behaviors. We infer that a person who can perform the sample of behaviors well can also perform the other, unmeasured behaviors well.

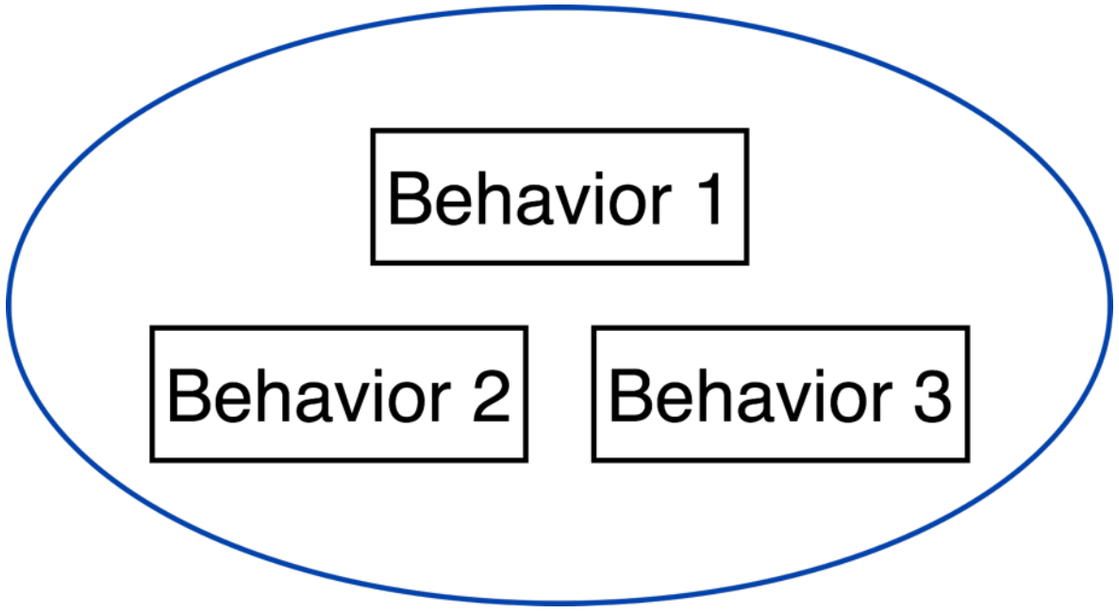
FIGURE 1 A Construct Is an Unobservable Cause for Observable Behaviors



Construct A causes Behaviors 1-3. Performance on the measured behaviors is used to infer standing on the unobserved cause for these behaviors.

So we have two definitions of construct. The first is the *unobservable cause* definition, and the second is the *set of behaviors* definition. Figure 1 and Figure 2 illustrate these two definitions. Both definitions are valid. Sorry about the confusion. I hear this double meaning thing doesn't happen in engineering. Finally, you should know that the follow-

FIGURE 2 A Construct Is a Set of Behaviors



Construct A is Behaviors 1-3. Performance on the measured behaviors is used to infer performance on the set of all relevant behaviors.

ing terms are often used as synonyms for construct: ability, dimension, trait, latent trait, factor.

Construct Standing: A person's status on the construct in question. If the construct is intelligence, then a person's standing might be high (i.e., smart), low, or somewhere in between.

Dimensionality: The number of constructs measured on a test. A test can be unidimensional (one construct only) or multidimensional (more than one construct). Life gets confusing when tests are multidimensional. Here's why. Consider the following four-question test (answered in a yes/no fashion) measuring the fictional constructs Delta and Zeta (because Greek letters are cool).

1. Do you like aspect 1 of concept Delta?
2. Do you like aspect 1 of concept Zeta?
3. Do you like aspect 2 of concept Delta?
4. Do you like aspect 2 of concept Zeta?

Now, let's say two people (Hermes and Nike) take this test. They both answer "yes" to two of the four questions. How do we interpret their scores? Did they score a two because they like Delta but not Zeta – or is it the other way around? Or do they like both Delta and Zeta but only in part? What we have here is a multidimensional test. We need to break it into two unidimensional tests,

one measuring attitudes toward Delta (Items 1 and 3) and another measuring attitudes toward Zeta (Items 2 and 4). Fortunately, we don't have to actually break the test in half. We can keep it as it is and just score it as two separate tests.

Important point: We're used to thinking of a test as what we can fit on a few sheets of paper, one test booklet equals one test. But the number of distinct tests (as indicated by the number of constructs measured) is determined by the number of ways we score the questions on the paper. One test booklet may contain as many tests as you like. To illustrate, when we score our incredibly interesting Delta/Zeta test two ways, we find that Hermes has a score of 2 on the Delta questions and a 0 on the Zeta questions. Nike has scores of 0 and 2, respectively. So now it's clear. Hermes likes Delta but doesn't like Zeta, and Nike is just the opposite. Our problem was solved by moving from a single multidimensional test to two unidimensional tests. Interpreting the meaning of a test score is far easier with unidimensional tests.

One of the things we'll see later is that we like unidimensional tests so much that we're willing to throw out items if removing them improves unidimensionality.

Measurement: Assigning symbols (usually numbers) to objects (usually people) so that the properties of the objects are accurately represented by the symbols.

The big question in psychological measurement is: Do the tests yield scores that accurately describe the properties of the object being measured? If you were to step on a scale and the scale said you weigh fourteen pounds, you would probably say that the scale is broken. This test (the scale) is not faithfully describing properties (weight) of the object (you). We're going to apply this same logic to other measurement devices to determine how well they are working.

Now let's put a few of these definitions together to address at an important concept. What is the purpose of measurement? If we're measuring people, we measure to determine a person's standing on the construct in question. And here's the kicker: We do this assuming (and hoping) that some people will have higher scores than others. That is, there should be differences among the scores for the people tested. A test that produces the same scores for everyone is a useless test. This concept is called discrimination. Note that this is not unfair discrimination. We definitely do not want the test to assign scores on the basis of irrelevant factors (like sex or race). By extension, any construct other than the desired construct is irrelevant. We want the scores to be a pure representation of the desired construct. And if people vary in their standing on the construct, then their scores on a test of that construct should vary as well.

Populations and Samples

The classic statistics class issues of populations, samples, and sampling error do not play a big role in psychometrics, but they are still relevant. Any measurement book without them would be incomplete. That said, we'll get by with just a brief treatment of them.

Population: Everyone relevant to a study. If your study is about people in general, then your population consists of every person on the planet. If your study is about students in an art history class being taught a certain way at a certain place, then your population is everyone in that class. Aside from studies with narrowly defined populations, we never measure the entire population. Sometimes researchers like to pretend that they have measured a population just because their sample is big, but they're just pretending.

Sample: A subset of the population. If there are ten million in the population, and you meas-

ure all but one, you've measured a sample. Samples can be small ($N = 23$) or large ($N = 10{,}823$). Smaller samples are likely to lead to greater error in our results. So we prefer larger samples. Bad news: Large samples are labor intensive.

Sampling Error

Sampling error is the difference between a statistic (e.g., mean) computed in a sample and the true population value for that statistic. As an example, let's say that we desire to investigate how well high school seniors know the capitals of the 50 states. Thus, the population consists of every high school senior (remember, the population isn't always everyone on the planet – it's everyone relevant to the study). It is obvious that it will be too much work to give our state capital test to every senior high school student. So via a random process we select 163 students and test them. And let's say that their mean score is 34 correct. Now that's a sample of people and their mean score represents our best estimate of the mean score for all the senior students. But this estimate is just that, an estimate, and it won't be perfect. Now for the sake of argument, imagine that we collected data from every single high school senior (i.e., the population). And the mean population score turns out to be 22 correct. That's not exactly a small difference between our sample value (34 correct) and the population value (22 correct). That difference is sampling error. And it's the price we pay for being lazy. Sometimes sampling error is big, or sometimes, by sheer luck, it works out to be zero for a given study. The rule to remember is this: Larger samples are likely to lead to smaller amounts of sampling error. So, we like large samples. The bigger, the better.

For the "larger samples lead to smaller amounts of sampling error" rule to work, everyone in the population must have an equal chance of being selected for the sample (such a sample is called a probability sample). There are a variety of

techniques (e.g., simple random sampling, cluster sampling) available to collect a probability sample. It's work, but it can be done. But what if the sample isn't a probability sample? The "larger samples lead to smaller amounts of sampling error" rule definitely does not apply if the sample is any type of non-probability sample (i.e., samples of convenience; volunteer samples; collecting data from friends, family, and pets). In a non-probability sample, some members of the population have no chance of being selected. The classic example of a non-probability sample is the use of college students in psychological research. Any sample taken from a college student subject pool will not be representative of any population broader in scope than college students for the simple reason that people who are not college students have zero chance of being selected.

Data gathered from a non-probability sample, regardless of size, should never be used to draw inferences regarding population characteristics; the validity of such generalizations is unknown and unknowable (Pedhazur & Schmelkin, 1991). No statistical magic exists which would fix the problems caused by the use of a non-probability sampling technique.

This next point should be obvious, but I'll state it anyway. The population from which the sample is taken must be the right kind of population. That is, it must be the population that is relevant to the study. Using our state capital example from earlier, if we wanted to know the average score of high school seniors, it wouldn't make sense to draw our sample from the membership of a plumber's union. If the sample is taken from Population A, we can't validly generalize sample characteristics to Population B.

To summarize matters, we can state a rule regarding inferences from samples. To make inferences from sample statistics to the population with a minimum of error, our sample (a) must be

large, (b) must be collected via a probability sampling technique, and (c) must be collected from the right type of population.

Finally, these rules apply to all statistics. We used means in our example, but we could have used medians, standard deviations, correlations, or half of a-thousand other statistics of which you have not yet dared to dream. Sampling error affects every statistic that we compute, and the only sure way to completely avoid it is to measure the entire population. Because that's too much work, we can minimize the magnitude of sampling error by the use of large probability samples.

Statistics and Statistics Related Accessories

2

Variability. Sampling Error. Standard Scores.

It doesn't get any better than this.

Introduction

Psychological measurement (or *psychometrics*, if you're cool) is not really about statistics. Various statistics are used and used often. But at the end of the day, we don't care that much about the statistics themselves. Statistics are just tools we use to evaluate various characteristics of our measurement. Furthermore, we will not dwell on the bane of every statistics class, significance testing. That is, in psychometrics there will not be an emphasis on ANOVAs, *t* tests, or the like. We will mostly live in the land of descriptive statistics: means, standard deviations, correlations, and regression equations. (The sample sizes in our psychometric projects are often so large that the concept of statistical significance is seldom a concern; Nunnally, 1967).

Levels of Measurement

Anyone can slap a number on something and call it measurement. That's fine, but we can't treat all of the numbers as if they have the same properties. The mere fact that one person has a score that is twice as large as yours (e.g., 100 versus 50) doesn't mean that his or her standing on the construct is twice as high as yours (i.e., if it is an intelligence test, they are not necessarily twice as smart). It may not even be the case that he or she has a higher standing on the construct than you. Thus, we must be careful that we do not infer more from the numbers than the type of measurement will allow. In the most extreme case, the numbers are just shorthand for names and are not meaningful in themselves. This level of measurement issue was analyzed by Stevens (1946). His system for understanding the various types of measurement is called Stevens's Scales.

CHART 1 Nominal Measurement

Characteristics	Numbers are just codes for categories. Greater numbers do not mean more of the construct
Example	Test taker gender (e.g., male = 0, female = 1)
Permissible Data Transformations	Anything that maintains the original categories (e.g, recode all males to -7; recode all females to +112)

The lowest (as in simplest) level of measurement is nominal, a word meaning *in name only*. Nominal measurement is summarized in Chart 1. It's measurement according to the definition, but it's not really measurement in the way we typically think of it – the numbers are completely arbitrary. The arbitrary nature of the numbers is why we can do so many kinds of transformations. Ordinal is where measurement starts to resemble the sort of measurement that we expected to see

CHART 2 Ordinal Measurement

Characteristics	Greater numbers indicate more of the construct but do not indicate how much more. A 10 point difference between scores might be big or small and isn't consistent across entire scale
Example	Class rank
Permissible Data Transformations	Anything that maintains the original order of scores (e.g., add 83 points to all scores)

(Chart 2). With ordinal data, bigger numbers do mean more of the construct. That's a big step. The big limitation in ordinal is that although bigger means more, we don't know how much more. A one point difference may be small, or it may be big. We don't know. Even if you did know that the difference between 53 and 54 is small, we don't know if the difference between 8 and 9 is equally small *even though both differences are one point*.

CHART 3 Interval Measurement

Characteristics	Greater numbers mean more of the construct, and the size of the difference between scores is meaningful. Zero doesn't mean anything special
Example	Many intelligence and personality tests
Permissible Data Transformations	Multiply all scores by a constant and/or add a constant to all scores (e.g., linear z score transformation)

Interval adds what we were missing from ordinal (Chart 3). With interval the size of the difference between points has a constant meaning. If a 10 point difference is a big difference, then it's a big difference at all points along the scale. That is, the difference between 60 and 50 is of the same magnitude as the difference between 35 and 25. The only thing we are missing in interval is a meaningful zero. In interval, zero is just another

CHART 4 Ratio Measurement

Characteristics	Everything that interval had plus a meaningful zero point. A zero on the test means zero of the construct
Example	Time to complete a task
Permissible Data Transformations	Multiply scores by a constant (e.g., transform time from minutes to seconds by multiplying by 60)

number. It does not mean the absence of the construct. A zero on an IQ test does not mean that the test taker is completely lacking in intelligence. Just to be weird we could set up the test so that zero is the highest score. It's silly, but we could do that. Wouldn't change a thing. Why is this zero thing so important? Because a meaningful zero is needed to say things like, "My score is 100. Yours is a 50. Therefore, I have twice as much _____ as you." (Fill in the blank with whatever construct the test is measuring.) We can't do that with tests

operating at the interval level. But with ratio we can. Ratio has it all (Chart 4). Bigger numbers mean more of the construct. The size of the difference between scores is constant and meaningful. Zero means the absence of the construct. I hope that you can see why ratio is mostly limited to physical constructs like size, weight, and time.

How do Stevens's scales impact the statistics we use? For nominal, we're pretty much limited to percentages (e.g., our sample was 53% male and 47% female), chi-square tests of association, modes, and the like. For ordinal, we have a bit more, but none of the really good stuff. If we want to compute the average of ordinal data, we're stuck with medians – not even a mean. We can do correlations, but not the regular kind. We have to use special ones like the Spearman rank-order correlation. It is not until we reach interval level measurement that we can use all of our favorite statistics. Thus, it is very important that we design our tests so that they have interval level measurement. How we do this will be covered on another day.

Here's an example using the latest results sent from my contacts down at the track.

Place	Horse Name	Number	Time
1	Seattle Tex	67	1:00
2	Four Sided Triangle	33	1:07
3	Mud King	92	1:21
4	Glue Factory Jailbreak	10	1:22
5	Mane Event	51	2:00

Nominal is exemplified by the number of the horse (e.g., Seattle Tex wears number 67). As should be clear, greater identification numbers do not indicate better performance. Ordinal is exemplified by the place the horse finished. First place (Seattle Tex) is the best. Second place (Four Sided Triangle) is second best. Mane Event is the worst. Ratio is shown by the time variable. Time indi-

cates the time elapsed from the start of the race until the finish. Mane Event took twice as long as Seattle Tex to finish the race (in other words, Seattle Tex was twice as fast as Mane Event). Where's interval, you ask? Sorry, I don't have one for this example; it doesn't lend itself to the world of horse racing. (Well, there's temperature in Fahrenheit or Celsius. That's arguably interval...)

Note how ordinal data doesn't indicate a consistent magnitude of the differences between scores. Mud King finished one place ahead of Glue Factory Jailbreak. Glue Factory Jailbreak finished one place ahead of Mane Event. It's a one place difference in both cases, but the size of the difference isn't the same. How do we know this? Look at the actual times (ratio level). Mud King finished one second ahead of Glue Factory Jailbreak. Glue Factory Jailbreak finished 38 seconds ahead of Mane Event. The one place difference doesn't have a constant meaning.

Distributional Statistics: Central Tendency

You've probably already learned that there are three types of averages: mean, median, and mode. An average score describes the central tendency of a set of data. The mode is the most frequently occurring value. Consider the data in the following table.

Person	Score
L. Sebastian	22
Kyle	18
S. Joe	29
Shauna	18
Ron	19

The modal score is 18 because it occurs more often (twice) than any other score (all just once each).

The median is the middle score. As an analogy, in a family with three children, who is the

middle child? The second one of course. If there are three scores, then the median is the value of the second score. So to compute a median, just find the middle score and obtain its value. In the above example, there are five scores so the middle score is the third highest one. The value of the third highest score is 19. Thus, the median score is 19 (not 3 or 29). It should be clear that to compute a median, one must (a) sort the data from highest to lowest (or lowest to highest), (b) find the middle score, and (c) obtain the value of the middle score. OK, new example. What if a family has four children, who is the middle child? It is a little tougher because two kids (the second and the third) tie for the middle spot. We could have the same issue with finding the median score. In the above dataset, let's say we obtain data from a sixth person, whom we will call Tammy. Tammy has a score of 20. That means we have six total scores. The middle scores are the third and fourth highest scores. Note that there are the same number of scores greater than and lesser than these two – that means that you've successfully found the middle value(s). The values of the two middle scores are Ron's 19 and Tammy's 20. To compute the median, split the difference. Thus, the median score is 19.5. To summarize, when we have an odd number of scores, just sort the data and find the value of the middle score. When we have an even number of scores, sort the data, find the values of the two middle scores, and split the difference.

You're probably most familiar with means. To compute a mean (symbolized as μ for populations and $\bar{X}$ for samples) add the scores and divide by the number of scores. If you like a good formula, here's one:

$$\bar{X} = \frac{\sum X}{N}$$

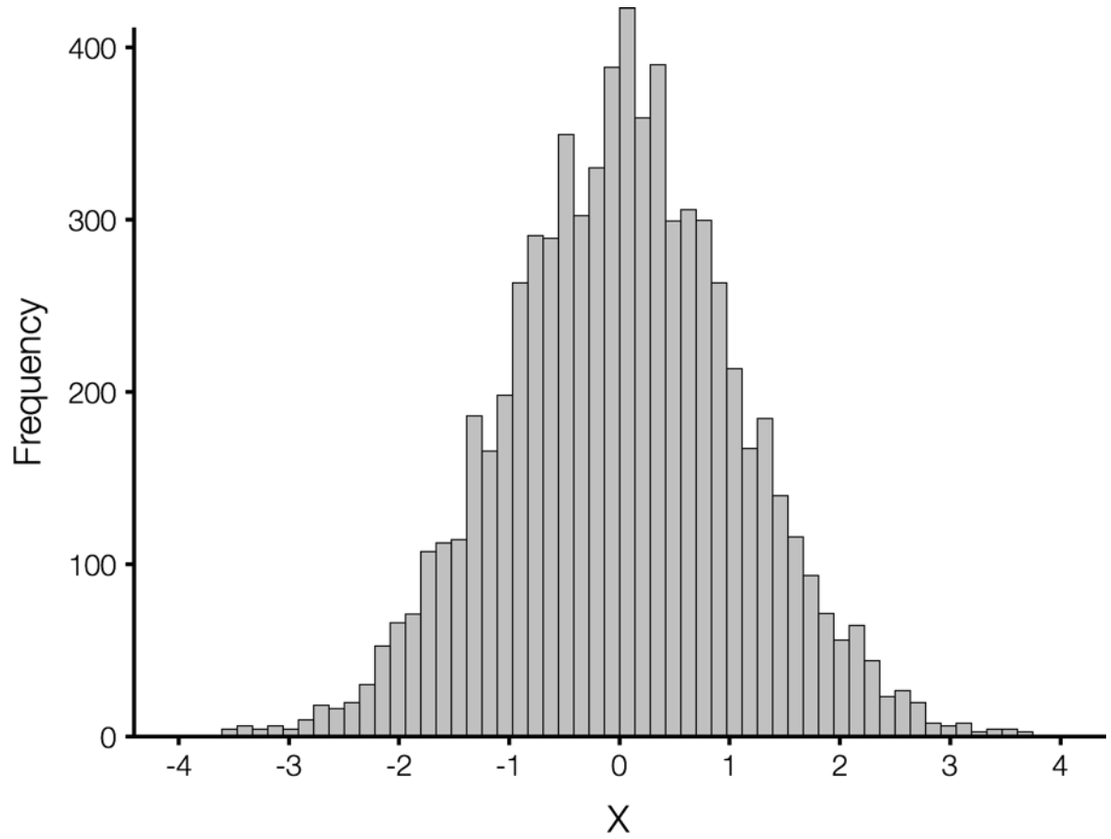
Now that you know three ways to compute average or central tendency, we should talk about the advantages and problems with each. There is

no problem with mode, except that nobody uses it. And I mean nobody. Means can be overly influenced by a single extreme score, resulting in a value that is not representative of the dataset. Medians do not suffer from that problem. In fact, one might say that medians are not influenced enough by extreme scores.

Distributional Statistics: Variability

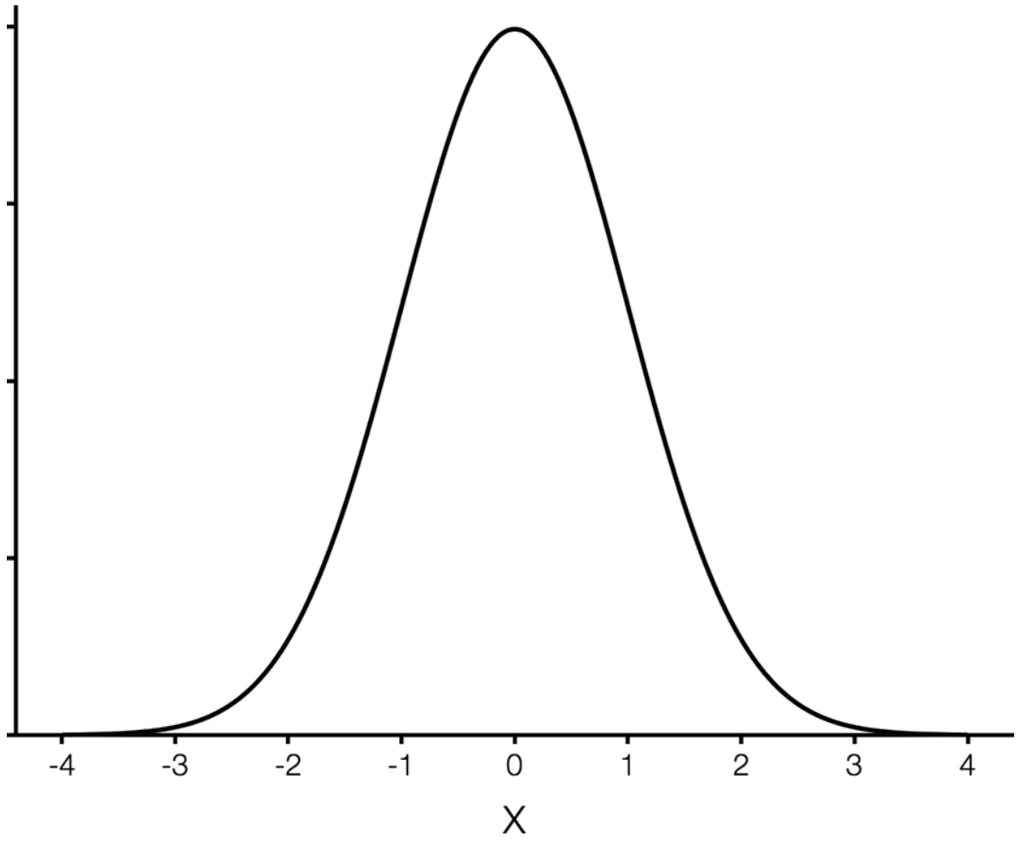
A frequency distribution (also called a histogram) is a graph of scores of a single variable (Figure 1). The *x*-axis indicates the various levels of the variable and the *y*-axis indicates the number of times each value is observed. It sounds fancy, but it's really just a bar graph, the sort of thing you made in third grade. The jagged nature of the bars is due to the *X* variable, a variable which has discrete categories (like ACT scores) where the only possible score values are integers (there's no 21.7). In other words, the variable is not truly continuous. With infinitely large datasets (and continuous variables), frequency distributions smooth out to something called a probability density function, (see Figure 2). Much nicer, no?

FIGURE 1 Frequency Distribution (Histogram)



Let us note a few of things in the two distributions. First, not many people have scores that are very low (-2 or -3) or very high (+2 or +3). Most people have scores in the middle ("The meaty part

FIGURE 2 Probability Density Function



of the normal curve." Costanza, 1997). Second, the Figure 2 distribution is symmetrical. If you draw a line down the middle, one side is a mirror image of the other. Go ahead, find a mirror and try it. We'll come back to this symmetry issue later. Finally, when the distribution is symmetrical like Figure 2, that line down the middle tells you where the mean is located. In this case, the mean is zero.

Moving on, distributions for two different datasets are displayed in Figure 3 and Figure 4. What's the difference between the two distributions? When they are shown on separate graphs they appear to be the same. They have the same mean score. Notice how the midpoint of each is zero. They have the same sample size (trust me on this). If you've read the title of this section, then you've guessed that the difference is variability. In the Figure 3 distribution (in black), most (approximately two-thirds) of the scores are within one point of the mean (the mean plus or minus one point), whereas in the Figure 4 distribution (in blue), very few of the scores are within one point of the mean. You have to move out to five points away from the mean (the mean plus or minus five points) in order capture most of the scores. If we place both datasets on the same scale (Figure 5), it's clear that the scores are not spread out in the same way (if Figure 5 seems like a massive cheat,

FIGURE 3 Variability Comparison: Low Variance

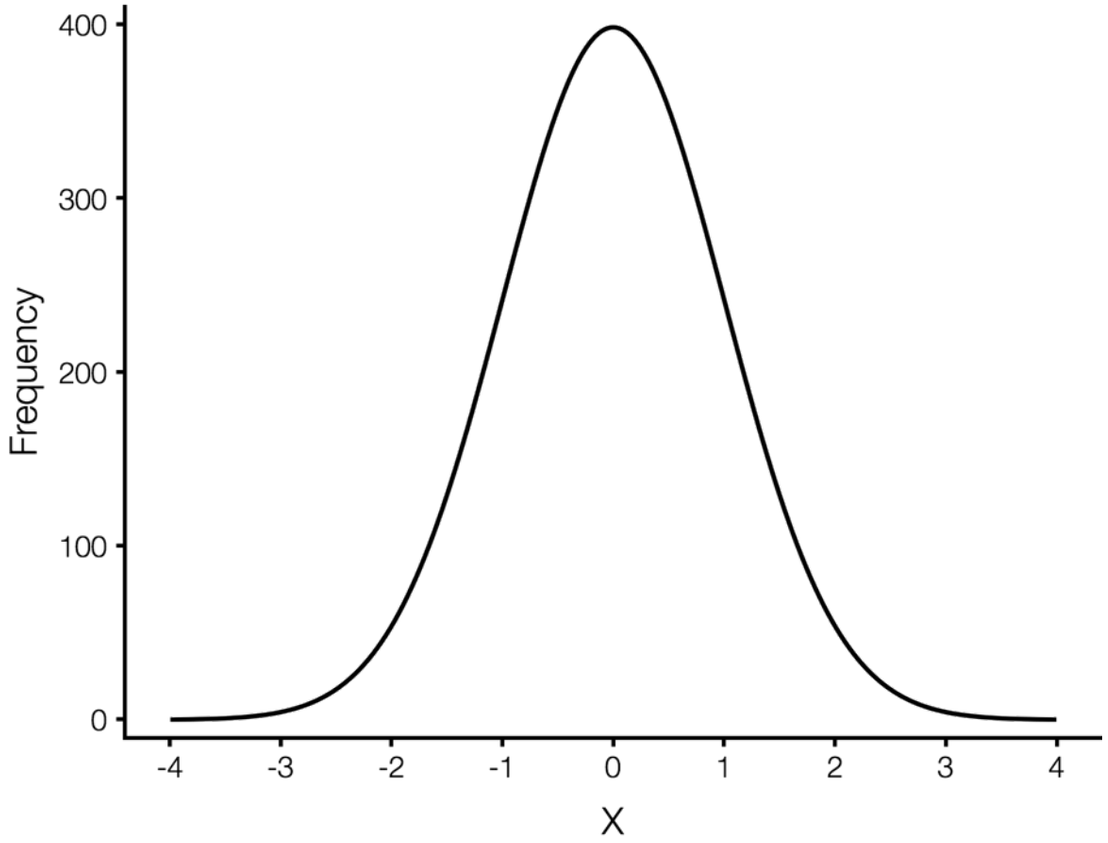
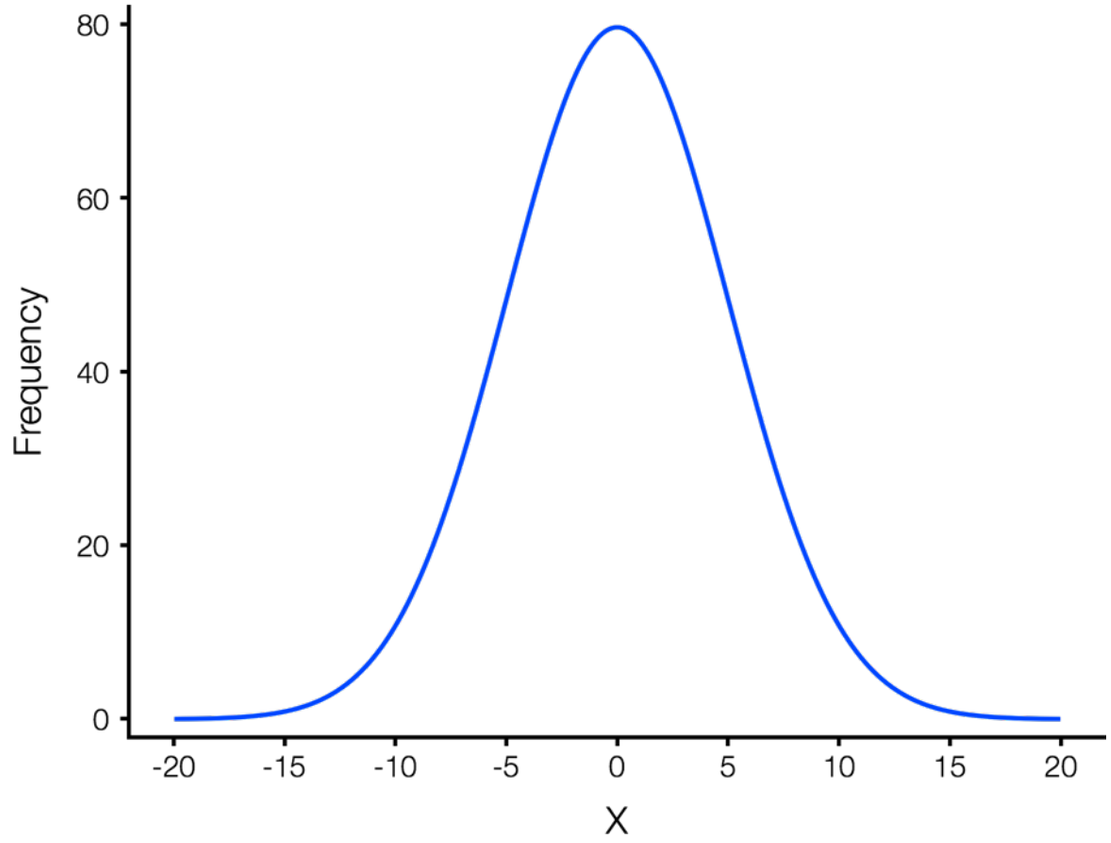


FIGURE 4 Variability Comparison: High Variance



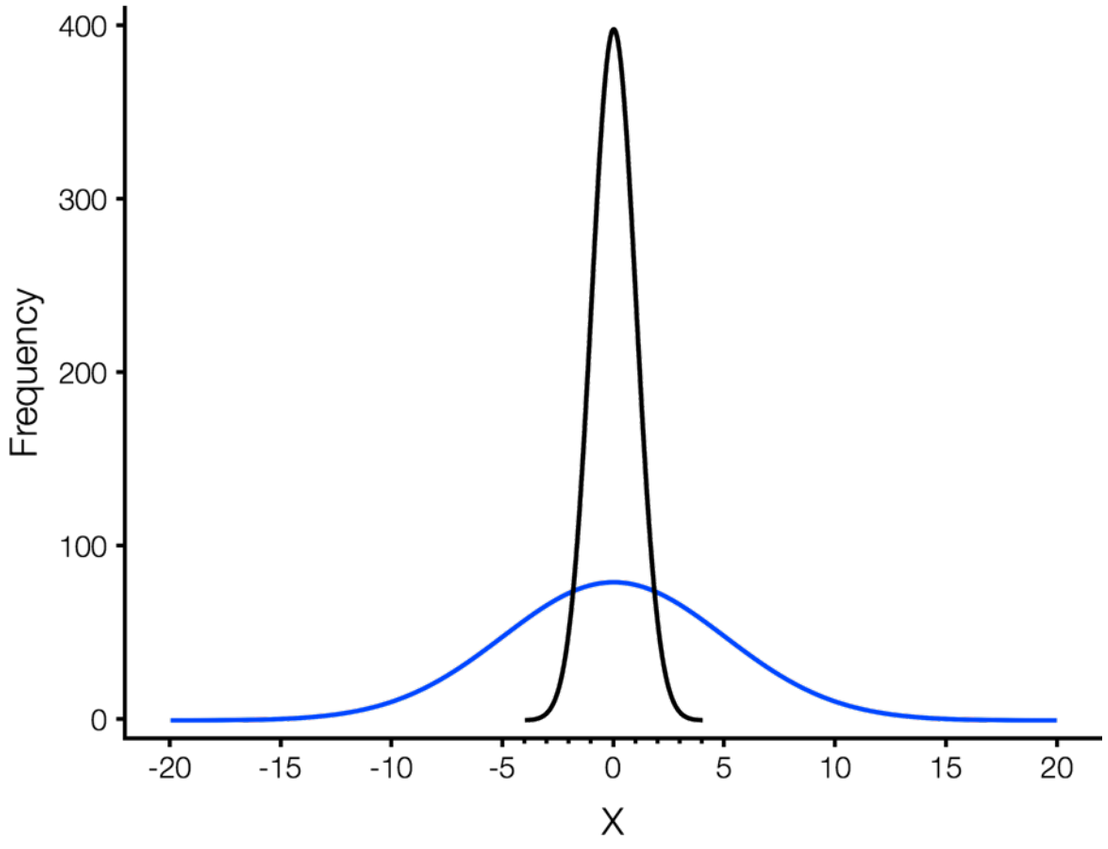
pay careful attention to the scale on the *x*- and *y*-axes on the three graphs).

Variability is greater for the blue distribution than for the black distribution. Variability is all about the differences between the scores. There are a number of ways to compute variability, but we'll end up using just two of them.

The simplest measure of variability is called range. The range is simply the difference between the highest and lowest scores. Easy to compute, sure, but range is a crude index of variability. A single outlying score can result in a high range.

A slightly more sophisticated measure of variability is the interquartile range. To compute the

FIGURE 5 Variance Comparison: Both Distributions



interquartile range, find the scores at the 75th and 25th percentiles (not unlike computing the median, which is the score at the 50th percentile) and compute the difference between the scores. The interquartile range is better than a simple range because it is more difficult for a single score to skew the results, but interquartile range is still not a sensitive measure of variability.

The measure of variability that we like is called variance (symbolized for populations as σ^2). Yes, the name is a little confusing, so here's a hint. *Variability* refers to all of these statistics (including range), whereas *variance* refers to a specific equation, given below for populations. (Just to be clear, the equation below computes variance for a population of data. But wait, you say, I thought people never measure an entire population. True. So why do we need this equation? To understand sample variance, we first need to understand population variance. All things in due time.)

$$\sigma_X^2 = \frac{\sum (X - \mu)^2}{N}$$

This equation isn't that bad. In fact, it is really similar to the equation for a mean. To see that, take all the parenthetical stuff and call it Q (just to give it a name). The equation is now $\frac{\sum Q}{N}$. Essentially, it is the mean of this Q variable. So variance is the mean of something. Now let's look at the

parenthetical component. It's $(X - \mu)^2$. Forget the squared part, focus on $(X - \mu)$. This is called a mean-deviation score and it is the difference between a score on X and the mean score. If X equals the mean score, then the mean-deviation score is zero. If X is greater than the mean score, then the mean-deviation score is positive. You get the idea. We'll be computing mean-deviation scores for all people in our dataset. An example is presented below. The mean of X is 6.

Person	X	(X - Mean)
Bennett	3	-3
Tommy	9	3
Todd	4	-2
Matt	8	2

Now to deal with the squared part, we'll simply square those mean-deviation scores.

Person	X	(X - Mean)	(X - Mean)^2
Bennett	3	-3	9
Tommy	9	3	9
Todd	4	-2	4
Matt	8	2	4

Remember that Q thing we made up? That's the last column, the squared mean-deviation scores. As we said, variance is just the mean of this thing.

So variance is the mean of the squared mean-deviation scores. In this case, it's $(9+9+4+4)/4 = 6.5$. Another way to describe it: variance represents the average squared difference between each score and the mean. Here's another example.

Person	X	(X - Mean)	(X - Mean)^2
Julianna	9	0	0
Paul	9	0	0
Jennifer	9	0	0
Anthony	9	0	0
Brenden	9	0	0

Variance is – you guessed it – zero. Why? Every score is the same. Thus, the average distance between each score and the mean is zero. Just for fun, diagram the frequency distribution of this dataset.

So that's the equation for population variance. What about the equation for computing variance when you measured a sample? (Which, as we have discussed, is pretty much all of the time.) The equation to compute the variance of a sample of data (when you want an unbiased estimate of the population variance – trust me, this is what you want) is:

$$S_X^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

The only difference (aside from the symbol S_X^2 and the replacement of μ with $\bar{X}$) is instead of dividing by N, we divide by $N - 1$. It is worth noting that the popular statistics programs (e.g., SPSS, SAS) use this $N - 1$ version for all of their variance computations (but Excel offers both equations – you pick the one you want). And, of course, the $N - 1$ version is the correct equation – unless you happened to have measured a population. And that won't happen on accident. So we'll stick with sample variance from here on out.

You might be wondering why we divide by $N - 1$ instead of N with the sample variance equation. Here's the short answer (and feel free to skip this paragraph if you don't care): the $N - 1$ denominator is necessary to obtain an unbiased estimate of the population variance. "Unbiased estimate?" you say. Well, think about it. We measure samples

because it's inconvenient (well nigh impossible) to measure the entire population. But, and this is important, we want our sample statistics to represent the population statistic. All of the statistics we have discussed to this point (e.g., mean) were unbiased, meaning that the sample statistic would not consistently yield a value that was too high or too low (stated another way, there was about a 50% chance that the sample statistic would be too high compared to the population value and about a 50% chance that it would be too low). Variance computed in a sample using the N denominator is a biased statistic in that it will consistently yield a value lower than the population value. And where does the $N - 1$ denominator come in? By dividing the squared mean-deviation scores by $N - 1$, the bias is eliminated, and the sample variance equation produces an unbiased estimate of the population value. Aren't you glad you asked? If you want to know why the N denominator version of the equation produces a biased estimate in a sample, that's a much bigger question. There are proofs for that. Take my word for it – they are not fun.

Our final variability statistic is called standard deviation (symbolized as S_X). If you know variance, then standard deviation is a snap because…

$$S_X = \sqrt{S_X^2}$$

That's right, standard deviation is just the square root of variance. If you know one, you can always compute the other. A clear sign of this is the symbol for each. The variance symbol (S_X^2) has a squared sign and the standard deviation symbol (S_X) doesn't.

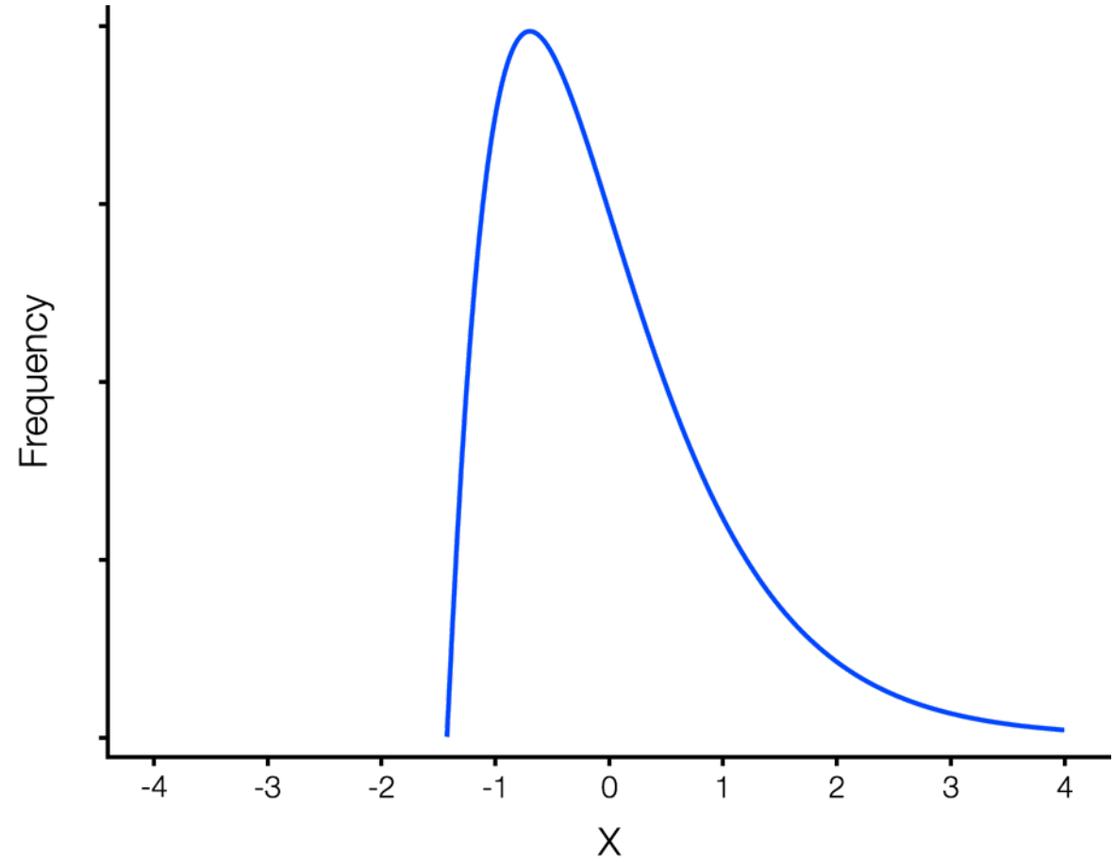
You might be tempted to ask, given that variance and standard deviation are basically the same, why do we need both of them? Well, that's a good question, and I'm glad you asked it. Reflects well on your intellect. The answer relates to the metric of measurement. If scores on X are the weight of people in pounds, and the variance

works out to 85, then we say the variance is 85 pounds *squared* because variance is in squared units. Right away you can see the problem: *squared* pounds. Now imagine that we measured ACT scores. *Squared ACT points?* Variance doesn't live in the land of regular units of measurement. But standard deviation does. With standard deviation, we're back to pounds, ACT points, and the like – the original metric of measurement. Operating in the original metric of measurement makes it a little easier to determine if a given value is big or small. In squared units, everything looks big.

Distributional Statistics: Skewness

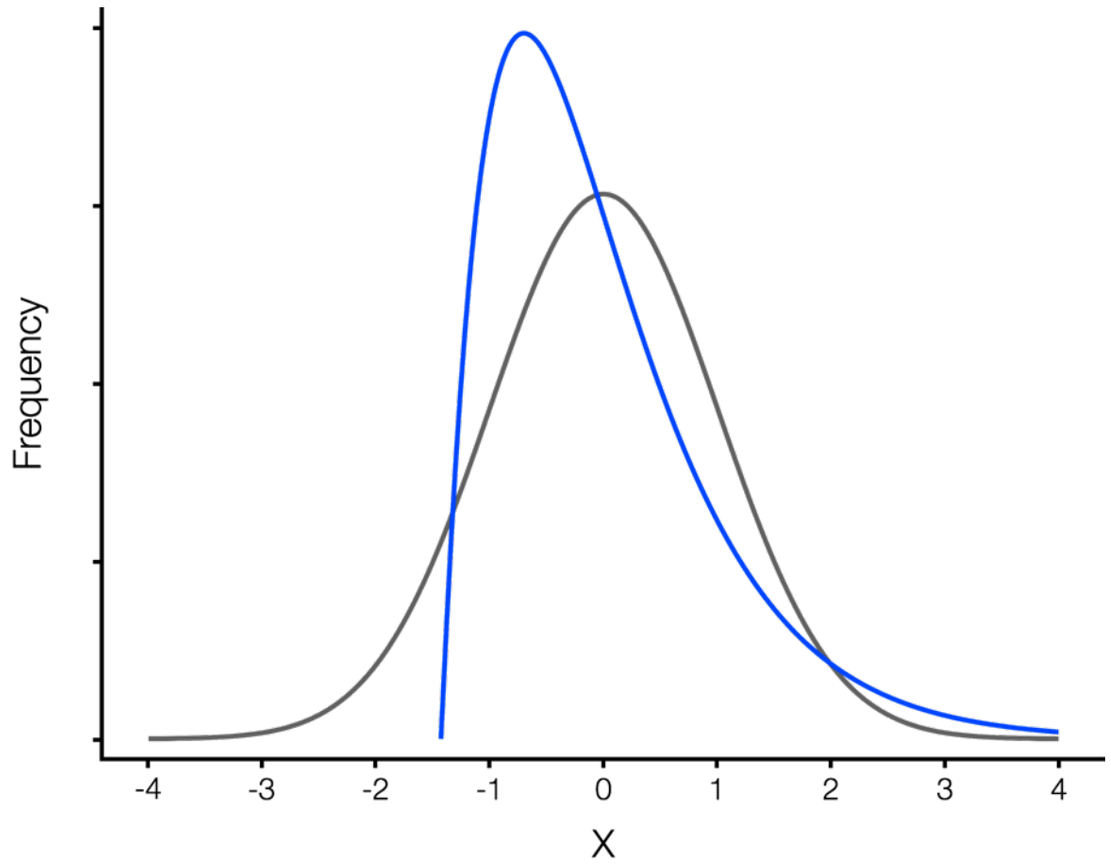
The distributions we have seen to this point have all been symmetrical; one half is a mirror image of the other half. Consider the distribution in Figure 6; it is not symmetrical. Would you believe me if I told you that Figure 6 has the same mean (0.0) and standard deviation as, say, Figure 3 (1.0)? It's true. I went to a lot of trouble to make

FIGURE 6 Skewed Distribution



it true. Same mean, same standard deviation but different shape. It does not have a symmetrical shape. We call it skewed. Figure 7 shows the skewed distribution (in blue) and a symmetrical distribution (in gray) on the same axes. Same means, same standard deviations, different shapes.

FIGURE 7 Skewness Versus Symmetry

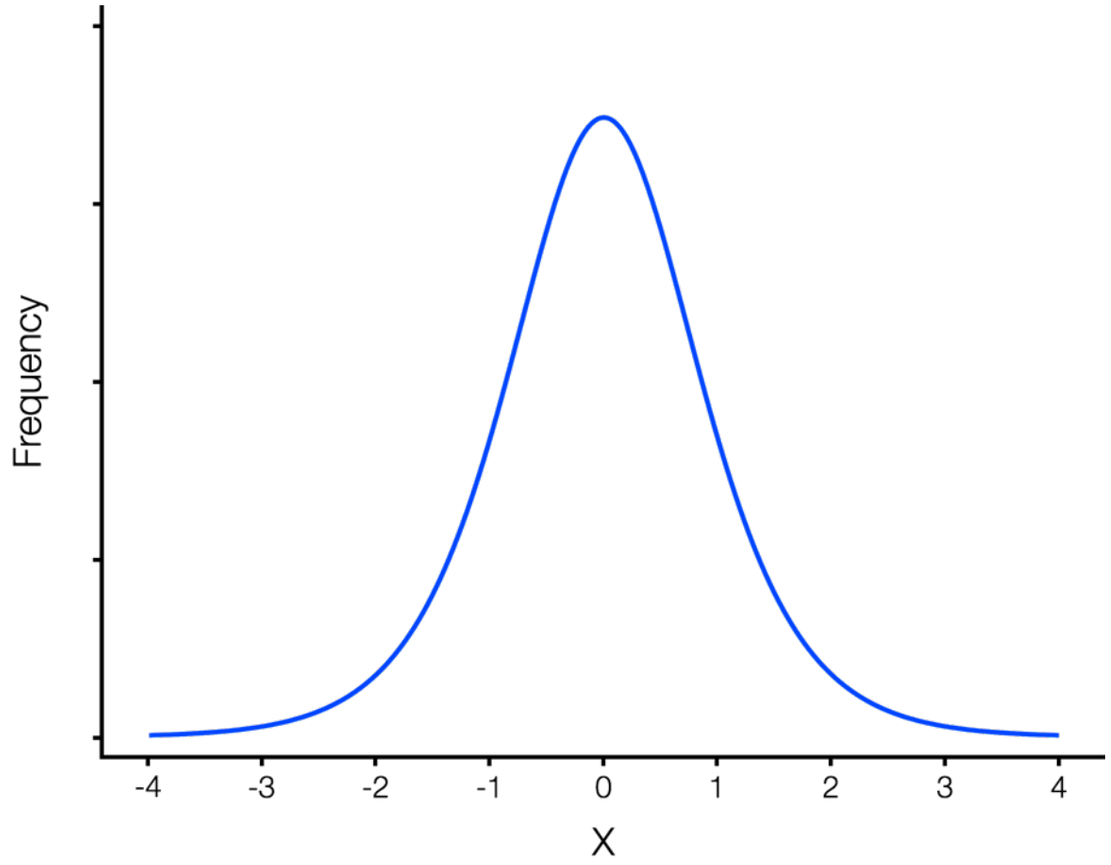


There is an equation to compute skew, but I won't burden you with it. If the value comes out to be zero, then there is no skew, meaning that the distribution of scores is perfectly symmetrical. If it's positive, then there are fewer scores at the high end (above the mean) than at the low end. If the skew is negative, then it is the opposite (too few scores below the mean). The blue distribution is a prime example of positive skew (skew = +1.4). And yes, I think the positive/negative labels are backwards too.

Distributional Statistics: Kurtosis

Examine the distribution in Figure 8. Looks pretty good, right? Bell-shaped. Perfectly symmetrical. Probably one of those *normal* distributions you hear people on the street talking about. But it's not. The shape is slightly off. You see, a normal distribution has a very specific shape. The normal distribution is not just any symmetrical distribution with most of the scores in the middle. In a normal distribution, a certain percentage of scores are at the midpoint versus the tails (the extreme ends) of the distribution. Now examine Figure 9. The distribution from Figure 8 is reproduced along with an actual normal distribution (in gray).

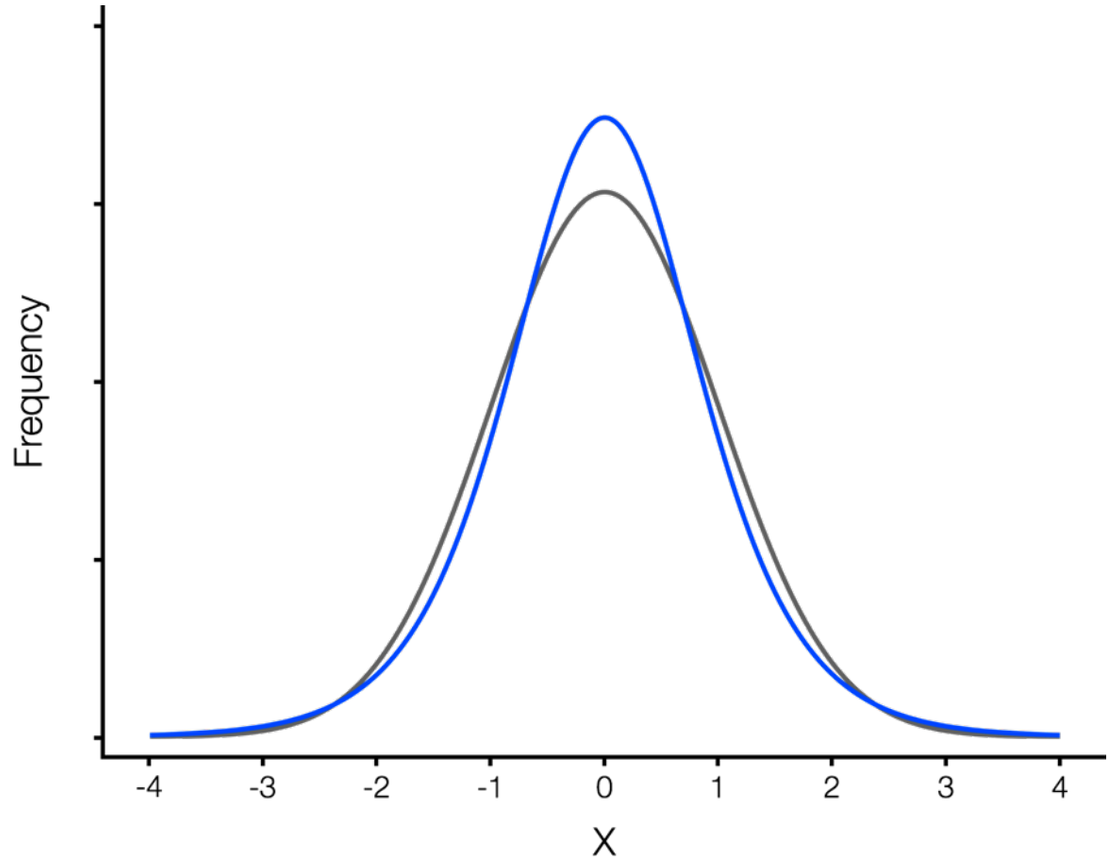
FIGURE 8 A Normal Distribution?



A nice distribution that looks like a normal distribution. *Looks like* being the key words

Both distributions have the same means (0.0) and standard deviations (1.0). They are both perfectly symmetrical. So what's the difference? The Figure 8 distribution has too many scores at the midpoint and not enough in the -1 to -2 and +1 to +2 areas as compared to a normal distribution.

FIGURE 9 Kurtosis Comparison



Distribution exhibiting non-normal kurtosis (blue) and a normal distribution (gray)

This issue is called kurtosis. Kurtosis is a function of the proportion of scores that are at the mean, close to mean, at the extremes of the distribution, and so on. A normal distribution has a kurtosis of zero.

Just to be clear, this isn't a difference in variability. Kurtosis sounds a little like variance, but it's different. Variance describes the average (squared) distance between each score and the mean. Kurtosis describes the proportion of scores that are close to and far from the mean. The above example is but one way for kurtosis to be off (i.e., not zero). As with skewness, I won't encumber you with a kurtosis formula. I will tell you that the kurtosis of the Figure 8 distribution is +1.2.

The Normal Distribution

Well, I spoiled the normal distribution in the previous section. I was trying to save it all for a big reveal here, but there isn't a good way to explain kurtosis without mentioning the normal distribution. So I blew the surprise. Sorry.

The normal distribution (or Gaussian distribution) has a bell shape, but not all bell shaped distributions are normal. All dogs are mammals, but not all mammals are dogs. A bell curve is called a bell curve because it kind of looks like the profile of a bell. The normal distribution is indeed bell shaped, so it is an example of a bell curve. There are many distributions that look bell shaped, but aren't normally distributed, as we saw in the kurtosis section. What then, is the difference between the true normal distribution and a mere bell curve? The normal distribution is a very specific shape. To be specific, the normal distribution has zero skew and zero kurtosis. You've seen enough normal distributions by this point, so I won't draw another one. (Almost every distribution in this chapter has been a normal distribution. The only distributions that weren't normal were the blue distributions in the sections on skewness and kurtosis.) The normal distribution can be described by the following equation (called a probability density function, and is exemplified in Figure 2), which I present for reference purposes only.

$$h_X = \frac{1}{e^{\left((X-\mu)^2 \big/ 2\sigma^2\right)}\sqrt{2\pi\sigma^2}}$$

Where:
h_X is the height of the normal curve at X.

Because normally distributed data always has the same shape, the normal distribution has many desirable properties, all related to zero skewness and zero kurtosis. First, because the normal distribution is symmetrical, we can talk about a score's relative position to the mean. That is, is a score above the mean or below the mean? How far above or below? Both questions are important and both have quantifiable answers that have the same meaning for all normally distributed data. (What if the data are not normally distributed? Well then, life's not so simple. We have the good fortune that most variables are approximately normally distributed.)

Here's what you'll find if you examine a set of normally distributed data. A certain percentage of scores will always be the same distance from the mean. For example, let's say you have a set of normally distributed data from a sample of 100 people. Let's also say the mean of this data is 0 and the standard deviation is 1. If you count how many people have scores between the mean and one standard deviation above the mean, you will find 34. If you count the number of people with scores between one and two standard deviations above the mean, you'll find 14. Finally, if you count the number of people with scores between two and three standard deviations above the mean, you find just two people. And because the distribution is symmetrical, things are the same for scores below the mean.

A few words about the preceding numbers. First, because I used a sample of 100 people, the numbers are rounded. Second, the previous paragraph also makes it appear that the normal distri-

bution goes no further than three standard deviations above or below the mean. The truth is that the normal distribution is without bounds; in theory, you could find someone with a score so high that they are seven standard deviations above the mean (or nine below the mean or whatever). These are scores so rare that we will not concern ourselves with them; we'll just focus on the world that is three standard deviations above and below the mean. It is this area that contains 99.7% of all scores. One last note, if a person's score is at the mean, their score is at the 50th percentile, meaning that it is higher than 50% of the scores. Figure 10 is a diagram of the normal distribution, divided into sections by standard deviation, showing the percentages in each section.

Now, what does all of this buy us? It allows us to quickly and easily attach meaning to a score. All you have to do is remember three numbers: 34, 14, and 2. If I told you that my score on a test was one standard deviation below the mean, what do we know about it? Obviously, it's below average. Using the 34/14/2 rule, we can estimate my percentile rank (the percent of people at or below a given score – we'll discuss percentile ranks in greater detail later). Now how do we figure this out? The only people with scores worse than mine are those with scores even lower than one standard deviation below the mean. A quick calcula-

FIGURE 10 Areas of the Normal Distribution

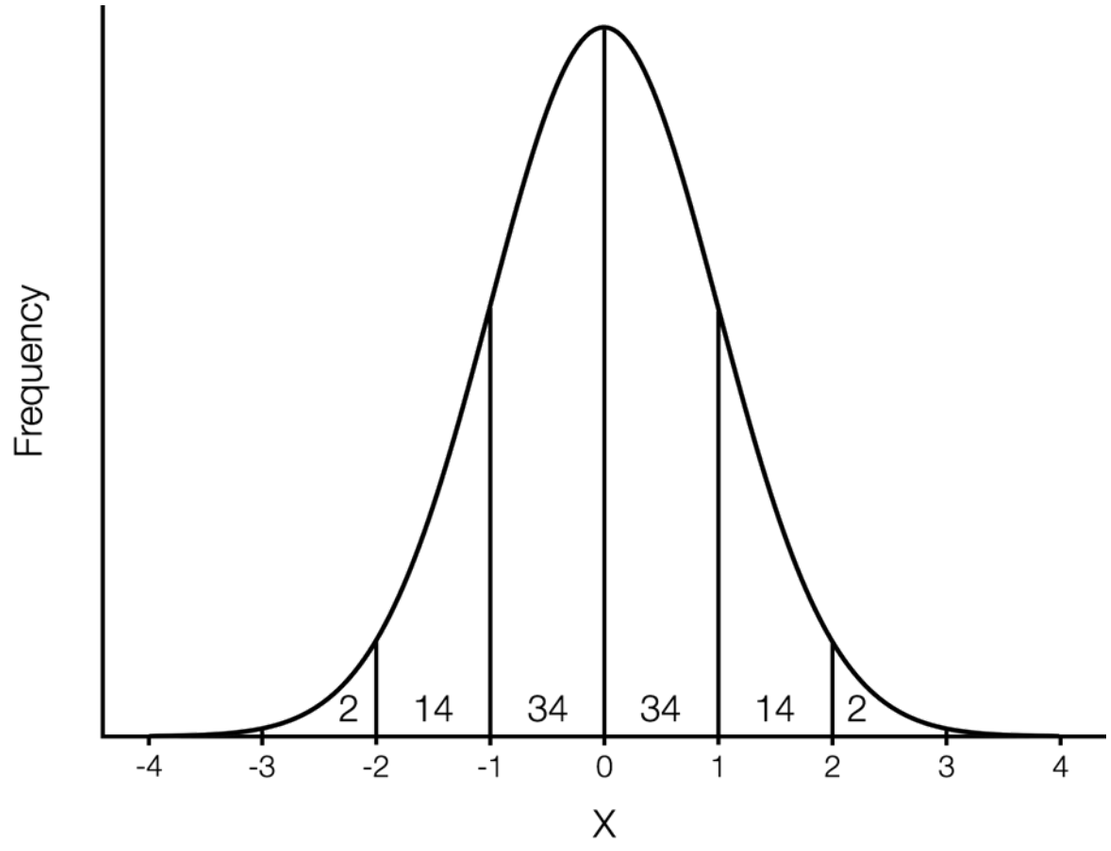
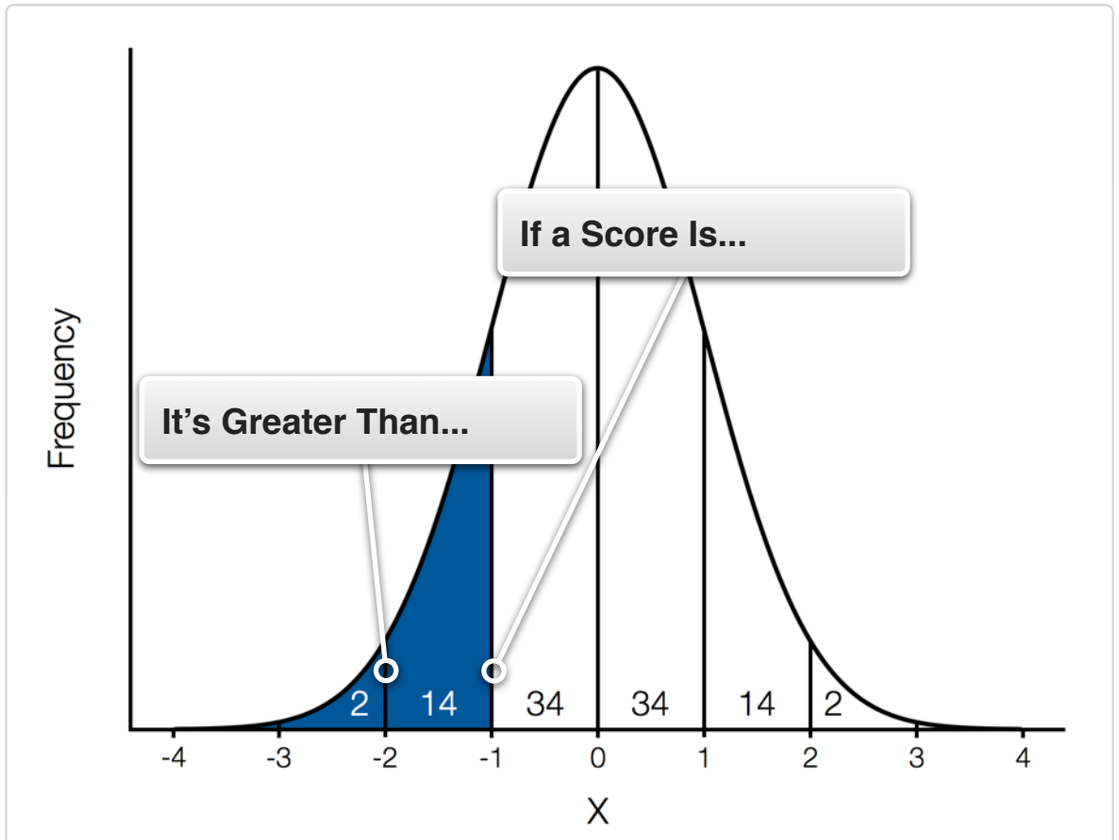


FIGURE 11 Using the 34/14/2 Rule to Estimate Percentile Ranks for a Normally Distributed Set of Data



tion shows that my score is greater than 2% + 14% of the scores. Thus, my percentile rank is 16%. This example is shown in Figure 11. So knowing the properties of the normal distribution helps us interpret test scores without much work. And it's all because normally distributed data always has the same shape.

I should stress one point that with this 34/14/2 rule for normal distributions: these percentile ranks are only crude estimates. If any precision is needed, consult a *z* table. Also, if the number of standard deviations above or below the mean for the score in question isn't a nice round number (e.g., 1.7 standard deviations above the mean), we'll need to consult a *z* table. And if the dataset isn't normally distributed, then forget 34/14/2 rule. And forget the *z* table. The z table is based on (and descriptive of) the normal distribution. Maybe this helps explain kurtosis. A dataset with a kurtosis other than zero will not have scores distributed in the 34/14/2 manner.

The only lingering question is this one: How do we know the number of standard deviations above or below the mean a score lies? As an example, if someone's score on a test is a 23, how do

we know the number of standard deviations above or below the mean? We'll have to compare that score to the mean score and use the standard deviation of the test to compute something. We'll save the rest of the answer to that question for a later section about something called linear *z* scores. Maybe you can figure it out yourself before we get there.

Normative Inference

So we just covered how easy it is to attach meaning to a test score with percentile ranks. But why do we need percentile ranks to interpret a test score? Here's the unpleasant truth: Scores on most tests have no inherent meaning. I can give you a measure of extroversion and tell you that you scored a 36. But what does a 36 mean? We don't know until we compare it to something. The most popular (but not the only) comparison is with other people's scores. If it turns out that your score of 36 is greater than almost everyone else's score, then I can say a score of 36 means that you are extremely extroverted. This process of giving meaning to a test score by comparing it to other scores is called normative inference. So that's the plain truth, most test scores have no inherent meaning. All we can say is, "Here's how you did compared to everyone else." Pay no attention to that man behind the curtain.

This process of normative inference is actually very familiar to you. You've been doing it your entire life in school. As an example, consider all of the times in class where you get a test back with a score that's really low, say 42. Based on the usual rules (90-100 = A, etc.), you have some idea of what a 42 means, but you need more information. So you ask questions like, "What was the average score?" "What was the highest score?" or "What's the curve?" If you find out that the highest score is a 72 and the average is a 65, you probably don't feel too good about your score of 42. (On the other hand, if you find out that the highest score

is 42, you feel great about your test.) That's normative inference in action. The test score has no meaning by itself, but it gains meaning when we compare it to other scores on the same test.

In order to make meaningful comparisons, the norm group, the group of people against whom we compare scores, must be representative of the population. (Side note: In some situations – academic tests in a single class – the norm group can consist of the entire population, but these are small populations.) Sampling issues were discussed in Chapter 1, so I'll just list the issues here. First, small samples are more likely to be affected by sampling error and, thus, are less likely to represent the population from which were are drawn. Second, for the sample to have a chance at being representative of the population, the sample must be drawn with a probability sampling technique.

What are the effects of using a norm group based on an unrepresentative sample? If I compare my score to a unrepresentative norm group, I'll conclude that my score is higher or lower than it really is. One way or the other, I drew the wrong conclusion.

One last issue. Consider your performance on the ACT. Who constitutes a relevant norm group? High school seniors or eighth grade students? It should be obvious that it's the high school seniors. We could compare your score to the scores taken from a large, probability sample of eighth grade students, but who cares? They're not relevant to our study; any comparison to them is meaningless. To summarize, if our norm group is to be a sample, we want it to be a large sample collected via a probability sampling technique from a relevant population of people.

Score Transformation Overview

You are already familiar with the way in which temperature values can be transformed from Fahrenheit to Celsius (or to Kelvin, if you're a fan of that one). A temperatures can be expressed in any of these formats with no loss in information. We often use whatever format is most convenient to us. You are also familiar with transforming raw scores (number of items answered correctly) on a test into a percent correct score (12 out of 20 becomes 60 percent). In both of these cases, you have transformed scores from one metric into another more convenient, or more useful, metric.

Score transformations are important in measurement for the simple reason that raw scores obtained from most tests are not all that useful in their raw score state. Thus, it is often to our advantage to transform these raw scores into another metric.

Standard Scores: Linear z Scores

Standardizing a set of data changes the scores so that they have a useful mean and standard deviation. We call these rescaled scores standard scores. There are many forms of standard scores. We'll discuss a few. Before we get to that, why would anyone use standard scores? As we mentioned in our section on normative inference, test score metrics (e.g., measuring race results in seconds versus hours, measuring job performance with a 5-point scale versus a 7-point scale) are arbitrary. Thus, it is difficult to interpret a score without knowing something about how other people score on the test. The mean and standard deviation are two pieces of information describing how well other people scored. Both statistics are used to transform raw scores into standard scores. Data expressed in standard scores allow us to interpret how high or low the score is as long as we know the characteristics of the standard scores.

Think of standard scores as a neutral playing field for our test scores.

There are many types of standard scores, but the most popular is the linear z score (often referred to as just *z score*, but the *linear* word is important as we will learn later). In fact, you already know a little about *z* scores based on what we learned earlier. The sample-based equation for computing a *z* score is very simple.

$$z_X = \frac{(X - \bar{X})}{S_X}$$

X represents the person's score in question. $\bar{X}$ is the mean score and S_X is the standard deviation. So, all we need to know in order standardize a score is: the test taker's score, the mean score, and the standard deviation.

How about an example? Let's say that I took the SAT, and my verbal score (SAT-V) is a 400. The mean of the SAT-V section is 500, and the standard deviation is 100. Now we're ready to go. Plugging in these values into the *z* score equation, we find that my 400 on the SAT-Verbal becomes a *z* score of -1.0.

Let's take a closer look at my *z* score of -1.0. My *z* score is negative. The negative sign tells you something – I did worse than average. If my score was above the mean, my *z* score would have been positive. If my score had been exactly the same as the mean, my *z* score would have been 0.0. The difference between my score of 400 and the mean is 100 points. The standard deviation is 100 points. Thus, my score of 400 is exactly one standard deviation below the mean. The *z* score is -1.0. Do you see where this is going? I'm not this redundant on accident. Here it comes: A z score is literally the number of standard deviations a score deviates from the mean. In case that's not clear, I'll restate the definition in the form of a question: How far (in terms of number of standard deviations) from the mean (above or below) is this score? If

the *z* score is -2.0, then the person's score is two standard deviations below the mean. If the *z* score is 1.5, then the person's score is one and a half standard deviations above the mean. If the *z* score is 0.0, then the person's score is zero standard deviations above the mean – it is right on the mean. So when we talk about the number of standard deviations a score is from the mean, we're also using *z* scores. Very convenient. The last thing to mention is that if our data are normally distributed, then we can quickly and easily attach meaning to the score with the 34/14/2 rule we learned earlier. Take my score of 400. In *z* score terms it is -1.0. If the data are normally distributed, that means my score is better than only 16% of the test takers (see Figure 11 again).

One important point about the linear *z* score transformation (and all other linear transformations) is that the shape of the distribution does not change. If the data were normally distributed before the transformation, it will be normally distributed after. It the data were skewed before, they will be skewed after. The linear *z* score transformation changes the mean and standard deviation of the data, not the shape of the distribution.

This is a good time to mention that there is a normal distribution called the standard normal distribution. What's the difference between the standard normal distribution and the normal distribution, you ask? Not much. In fact, the only difference is that the standard normal distribution has a mean of zero and a standard deviation of one. Thus, it's a normal distribution in *z* score terms.

(If you want the equation for the height of the normal curve at *X* for a standard normal distribution, it's the same as before, only with 0 substituted for μ and 1 substituted for σ. Making those substitutions simplifies the equation to:

$$h_X = \frac{1}{e^{(.5X^2)}\sqrt{2\pi}}$$

Not that you couldn't have figured out how to put a zero and a one in the original equation yourself. I just like the way the simplified version looks.)

So, to summarize things this far. Linear *z* scores indicate the number of standard deviations that a score lies above or below the mean. They are incredibly useful for interpreting test scores. Because the transformation of raw scores into *z* scores involves comparing the raw score to the mean and standard deviation of a group of people, they allow for easy normative inferences. If the data are normally distributed, then we can make a quick estimation of the *z* score's percentile rank by using the 34/14/2 rule. More precise estimates require a *z* table.

There's another benefit to standard scores. Standard scores allow for easy comparisons of scores. Comparing two or more scores from the same test is child's play if the measurement is done at the ordinal level or better – the highest score represents the highest standing on the construct. Highest number wins. But what if we want to compare scores from one test to scores from a different test? This won't be as easy. Now you might ask, why would anyone want to do this? The answer is that we have many similar tests that do the same thing. The ACT and the SAT offer but one example. Let's say that you took the ACT and scored a 30. We already know that I took the SAT and scored a 400 on the verbal section. Who did better, me or you? A layperson might look at the scores and say that I did better because 400 is bigger than 30. But we know better. We know that each test has a different metric of measurement – they use different numbers with different standards for good, average, and poor performance. What we need is a way to put both scores on the same metric of measurement. All we have to do is translate both scores to standard scores.

Back to our ACT-SAT example. We know my 400 on the SAT-Verbal translates to a *z* score of

-1.0. What about your 30 on the ACT? What's its *z* score? Using the *z* score equation (we'll say that the ACT has a mean of 20 and a standard deviation of 5), your ACT score transforms to a *z* score of +2.0. Again, *z* scores are the number of standard deviations above or below the mean. So your score is two standard deviations above the mean. If ACT scores are normally distributed, then you did better than 98% of the test takers. Very nice. Now that both of our scores are in *z* score units, we can directly compare the numbers. It is clear that your *z* score of 2.0 is bigger than my *z* score of -1.0. You win. You did better on your test than I did on mine. Try to stay humble. It won't be easy.

One last bit on this comparison business and then we'll move on. Some comparisons are not meaningful. Suppose you take a test of depression and I take the SAT-Verbal again. Your score is a 4 and mine is a 410 (I studied a bit harder this time). Who did better? The answer is: Who cares? The tests are completely different, measuring dif-

REVIEW 1 *z* Scores

Question 1 of 2

If a set of data has a mean of 50 and a standard deviation of 20, what is the *z* score for a person with a raw score of 40?

- ✓ **A.** -0.5
- **B.** +0.5
- **C.** +2.0
- **D.** -2.0
- **E.** +0.75
- **F.** -0.75

◀ Check Answer ▶

ferent constructs, existing for different purposes. It's a meaningless comparison.

Other Linear Standard Scores

Hard as it may be to believe, not everyone loves *z* scores. Probably the second most popular standard score is the T score. *T* scores are similar to *z* scores except that *T* scores are set to have a mean of 50 and a standard deviation of 10. Knowing this, it is easy to transform *z* scores to *T* scores (and back again). If my *z* score is 0.0, what's my *T* score? Answer: 50. Why? My *z* score is 0.0 which means that I am zero standard deviations above the mean. *T* score are set to have a mean of 50 and my score is right at the mean, thus it's 50. New example. My score in *z* score units is -1.0. What's my *T* score? My *z* score of -1.0 tells us that I am one standard deviation below the mean. So in *T* score-land, I start at the mean of 50 and go down one standard deviation (10 points) to 40. My score in *T* score units is 40. Last example. My *z* score is 3.0. What's my *T* score? It is 80, which is three standard deviations above the mean.

Now let's score the other way. My score in *T* score units is 35. What's my *z* score? This is easy. Recall that *T* scores are set to have a mean of 50 and standard deviation of 10. Just use the *z* score equation: (35-50)/10 = -1.5. My *z* score is -1.5 meaning that my score is one and a half standard deviations below the mean.

Why would anyone want to use *T* scores when *T* scores do the same thing that *z* scores do, only not quite as elegantly? This is just my opinion, but I think there are two answers, neither of which are very compelling. First, with *z* scores half of the data will be negative. Scoring in the negative range sounds like your score was terrible, but in *z*-score-land, it's only below average. So, for the protection of test taker egos, we may want to report scores that are positive. With *T* scores, it's very difficult to get a negative score. (Just how low would

you have to score to have a negative *T* score? There's something to think about the next time you're stuck in traffic.) The second reason is related to what we're used to in school. We've been used to the 90-100 is an A, 80-90 is a B, etc. system for many years. Thus, we're used to the idea that a 10 point difference is a big difference and a one point difference is a small one. (Important note: A one standard deviation difference is huge.) Now, in *z* score terms, a one standard deviation difference is one point, but in *T* score terms it is 10 points. Thus, *T* scores conform to our idea of a big difference in scores. A ten point *T* score difference looks big, but a one point *z* scores difference looks small – even though they are really the same size. Again, this is what it looks like to people unfamiliar with statistics. We know better. A one standard deviation difference is big. The difference between *z* scores of 1.4 versus 2.4 is big. Even if it is just one point.

It should be clear at this point that we could create any new standard score system we want and be able to transform our data to it. Let's say that we develop our own standard score system. We'll call it..., let's see, X, Y, Z are already taken. What's after Z? Omega? We'll go with that. Omega-Scores. Omega-Scores are designed to have a mean of 27 and a standard deviation of 11. If your *z* score is +2.0, what is your score in Omega-Scores? It's two standard deviations above the mean: $2 \times 11 = 22$. Thus, it's 22 points above the mean. 22+27 = 49. Let's say someone else has a *z* score of -1. Their Omega-Score would be 16. Why 16? The mean is 27 and they were one standard deviation (11) worse than the mean. Can we start with Omega-Scores and end up with *z* scores? Sure, just use the *z* score equation. Let's say our Omega-Score is 5. The *z* score would be (5-27)/11, which equals -2. The key to all of this is to remember: a) the *z* score equation and b)

that z scores tell you the number of standard deviations above or below the mean.

Percentile Ranks

Standard scores aren't the only ways in which we can change our data. Another popular transformation is the percentile rank, a change with which we are already familiar. Let's be clear about this: it is perfectly fine to compute someone's percentile rank in order to give more meaning to their test score, but it is not a good idea to throw away the raw score in the process. What I mean is, consider the percentile rank to be a supplement to the raw score, not a replacement for it. Why? Percentile ranks are ordinal data, which means that we can't do much (statistically) with them. They help us attach meaning to a test score, but that's about it. If we want to perform all of our favorite statistical operations, we need the original raw scores.

How do we convert raw scores to percentiles? If the data are normally distributed, we can consult a z table. Or we can just do it by hand – something that works for all data, regardless of the shape of their distribution. How? Here goes… First, sort the scores from highest to lowest. Second, count how many people have scores below the score in question (we'll call this $N_{<X}$, the number of scores less than X). Third, count how many people have the score in question; there may just be one person or a bunch of people tied at this score (we'll call this N_X). Fourth, count the total number of people (N). Finally, plug into the equation below.

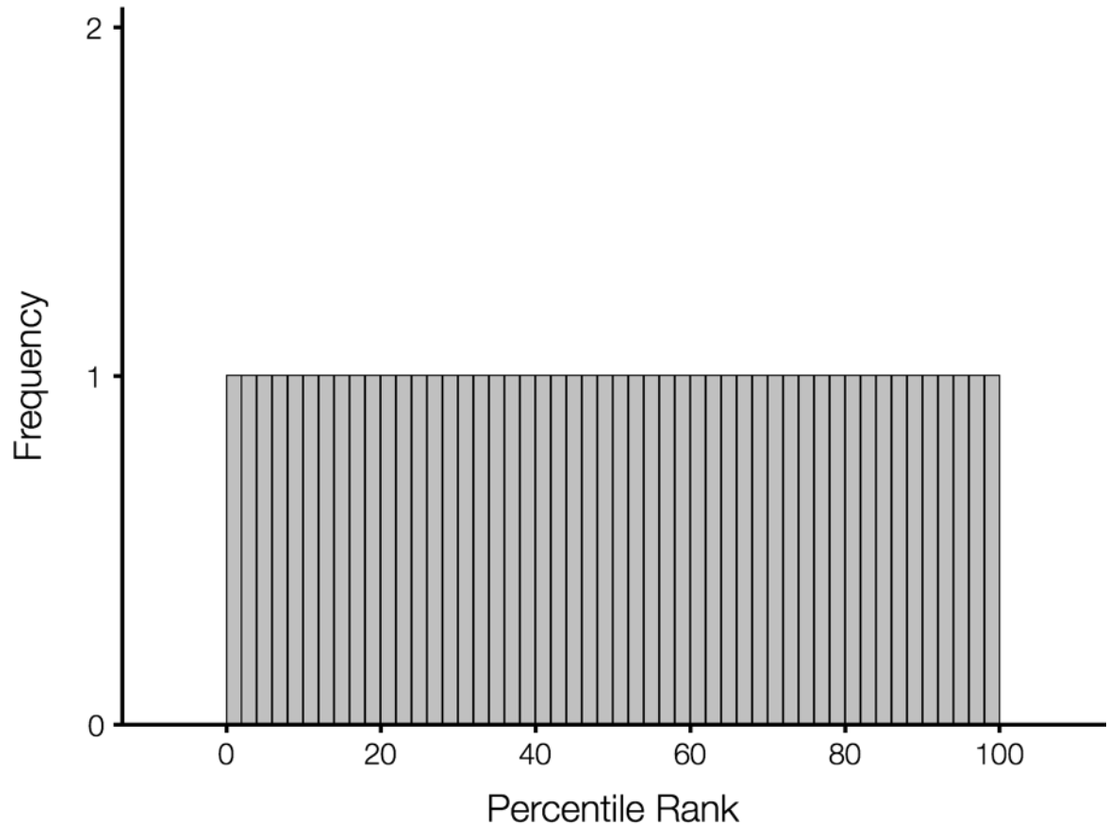
$$PR = \left(\frac{N_{<X} + (.5)N_X}{N} \right) 100\,\%$$

Where:
PR is percentile rank.

Last note on percentile rank transformations. After the transformation, what is the shape of the

distribution? Let's say it was normal before we started (i.e., the raw scores were normally distributed). After a transformation to percentiles, the distribution will be rectangular. As you can see in Figure 12, each percentile rank occurs exactly one time (aside from rounding issues). It should be obvious why we don't want to throw out the raw scores. (By the way, I'm sure I'll get no argument when I say that the rectangular distribution is least interesting distribution in the history of, well, distributions.)

FIGURE 12 Percentile Rank Distribution



Normalized z Scores

Another type of *z* score is the normalized z score. It sounds like a regular, linear *z* score, but with one huge difference, the *normalized* part. When we normalize something, we change the shape of the distribution to force it to be normally distributed. If the data were already normally distributed, then nothing changes but the mean and standard deviation, just like a linear *z* score transformation. But, if the distribution was say, seriously skewed, converting the raw scores to normalized *z* scores changes the shape of the distribution to perfectly normal. For an example of the changes that happen with normalizing, look back to Figure 7 or Figure 9; the blue distribution represents before normalizing and the gray distribution represents after, all nice and normal. Think of those dis-

tributions as one of those weight loss ads you see with before and after pictures. Only with normalizing you're not losing weight – you're losing skew and irregular kurtosis.

As you can see, the raw scores are positively skewed. After the transformation, the normalized scores are normally distributed. This change is non-linear, which means that the scores are not changed by a constant amount. The change occurs by compressing the scores at the high end (the difference between high scores is reduced; a two point difference before the transformation becomes a one point difference after the transformation) and stretching the scores at the low end (a one point difference becomes a two point difference). Scores in the middle may not be changed by much. Compare this with any linear transformation (including linear *z* scores) in which all scores are changed by a constant amount, regardless of whether they are high, medium, or low (e.g., if we measure time in minutes, but decide to change the scores to time in seconds, we multiply everyone's score by 60; a one unit difference now becomes a 60 unit difference across the board). This normalizing business is officially a big deal. It should be clear that normalizing should not be undertaken without good reason. What's a good reason? I've never seen one.

How are normalized *z* scores computed? Given that I more or less just told you that you shouldn't do it, it seems strange to describe the steps for doing it. But here goes. Normalizing is a two step process. First, transform raw scores into percentiles. Second, use a *z* table to find the *z* score associated with each percentile rank. That's it. It's simple and lethal. No more distributional problems. Everything is normally distributed.

Let me close with a comment on normalizing. The idea that normalizing is a valid technique we can use to solve our problems regarding normality assumptions is a relic from an outdated mentality

regarding data. The prevailing philosophy of the time, a simpler time, was that normalizing (or some other non-linear transformation) can solve your distributional problems. This line of thinking is almost always incorrect. The correct approach is to analyze the data with the correct statistics given the properties of the data. To be clear on proper procedures, don't perform some non-linear change to the data; do change the statistics you use to analyze the data.

Correlation and Regression

If loving regression is wrong, I don't want to be right.

3

Overview

You may have noticed that everything we've discussed so far has been related to scores on a single variable. That is, we've talked about a set of ACT scores, but we've never looked at the relationship between two variables (ACT and college GPA, just to throw out a wild idea) for a group of people who each have scores *on both variables*. Are the scores related? Unrelated? In what way are they related? Is it a strong relationship or a weak one? As you can see, life gets much more interesting when we measure multiple variables for each person. And we haven't even talked about *why* these two variables are related. That's a topic for another day (and another book). For now, let's focus on understanding how we quantify associations between two variables.

Bivariate Associations

When describing the association between two variables there are two issues to consider: the strength of the relationship and the direction of the relationship. One way to assess the association between two variables is to simply examine the raw data. Below is another one of our absurdly small datasets, which we'll use as an example.

Person	X (ACT)	Y (GPA)
John	12	1.1
Sal	23	2.8
Tim	24	2.9
Amy	31	3.4
Antonio	10	-

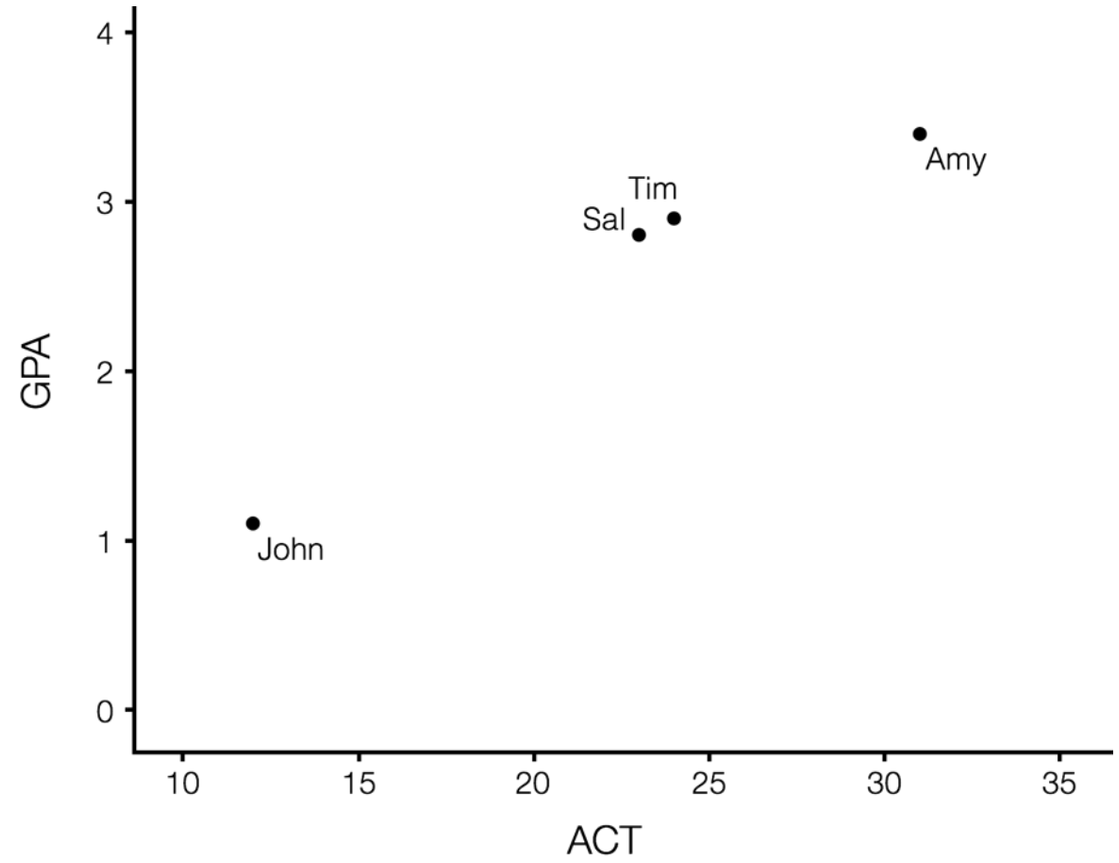
First off, we note that each person should have two scores: *X*, the ACT score, and *Y*, the GPA. If a person had only one score, we would be unable to include him or her in the analysis. Note that Anto-

nio doesn't have a score on the *Y* variable – we are unable to include him in the analysis. A person must have scores on both variables to be included. We also note that I've sorted the scores from lowest (John) to highest (Amy) on *X*. Now let's see if there's a trend in the data. And because I made up the data, there is. Lower scores on *X* are associated with lower scores on *Y*. Higher scores on *X* are associated with higher scores on *Y*. So it appears that there is a strong, positive relationship between *X* (ACT score) and *Y* (GPA). We say that it is a *strong* relationship because the rank-order is perfectly consistent. The person with the highest score on *X* (Amy) also has the highest score on *Y*. The person with the second highest score on *X* (Tim) also has the second highest score on *Y*. And so on. There are no exceptions to this perfect ordering of the scores. This consistency of rank-order is the primary determinant of the value of the correlation coefficient. Finally, we say the relationship is *positive* because higher scores on *X* are associated with higher scores on *Y*. If higher scores on *X* were associated with lower scores on *Y*, then the relationship would have been negative. Thus, we have addressed both aspects of bivariate associations: strength (the relationship between *X* and *Y* is strong) and direction (the relationship is positive). Now this is about all we can get from examining the raw data (don't try doing even this with large datasets – it's borderline impossible); let's move on to a better way to examine the relationship, the scatterplot.

The Scatterplot

A scatterplot is a graph of the *X* and *Y* scores on two axes. It's the same old kind of *x*-*y* graph you've known since, oh, about third grade. The data from our example are graphed in Figure 1. On a scatterplot, each person receives a dot (or a square, or a plus sign, or a smiley face, or whatever you want). The dot indicates a person's score on *X* and *Y*. It should now be clear as to why we

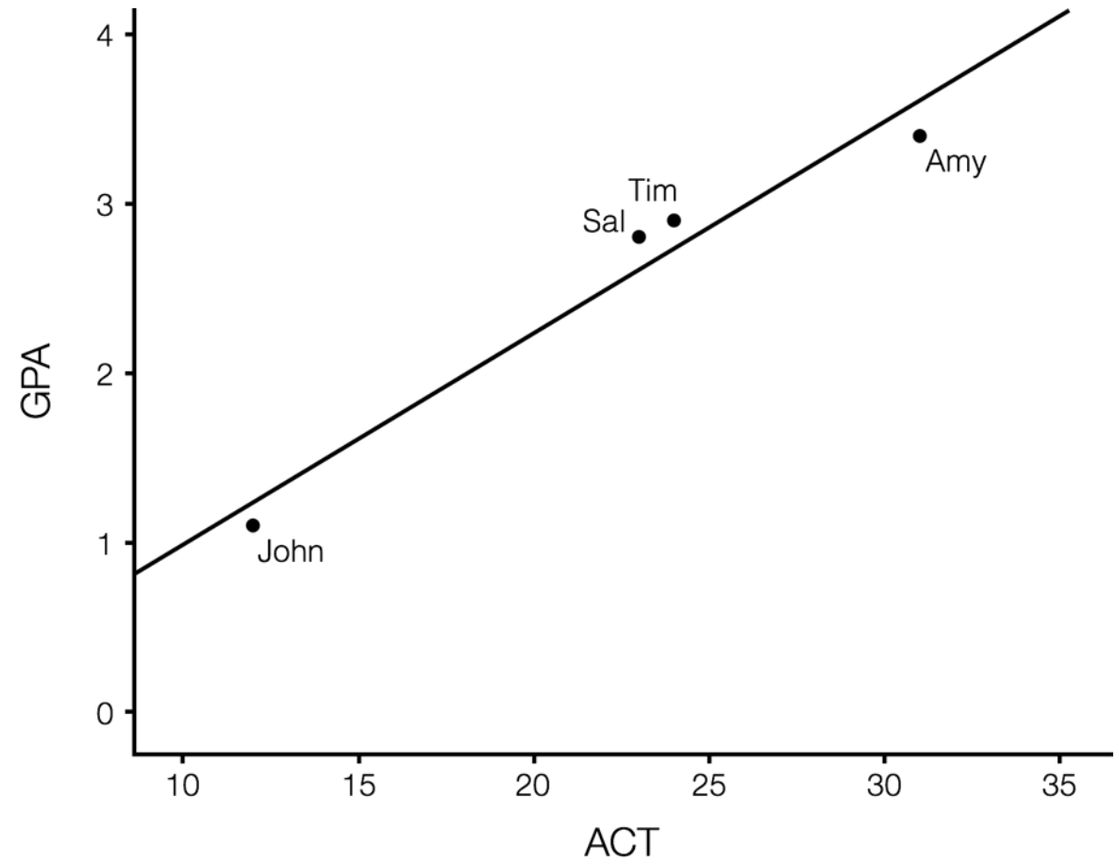
FIGURE 1 Scatterplot of Example Data



couldn't include Antonio in the analysis. Where would we put his dot? It would be somewhere at 3.0 on the *x*-axis, but how high on $X = 3$ do we put the dot (*y*-axis location)? We can't assume that he would have done poorly on *Y*. We're not in the business of assuming anything – we're in the business of using the available data to describe the relationship. Thus, he's gone. Looking at the scatterplot, we can see a trend, the same trend we saw when we looked at the raw data: higher scores on *X* are associated with higher scores on *Y*. And notice how the scores fall in the path of a straight line. The basic Pearson correlation tells us the strength of the *linear* relationship between two variables. What if the relationship is not linear? Another time, another book for that topic.

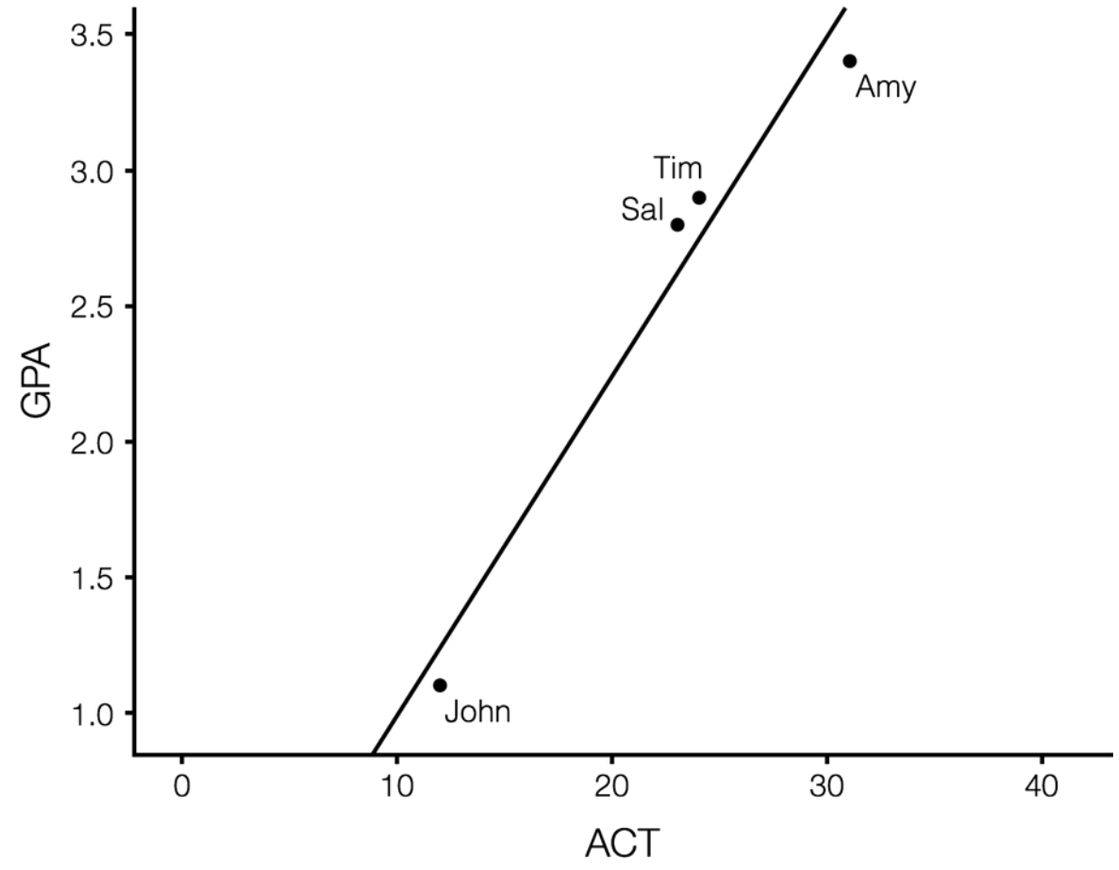
Getting back to how closely the scores match a straight line, let's draw the graph again, only this time with a straight line added as a reference (Figure 2).This line is called the *line of best fit*, or more commonly, the regression line. The regression line is the line that minimizes the vertical distance between the line and each point. You can imagine pulling out a ruler, measuring the vertical distance between each point and the line, averaging the distance, moving the line ever so slightly to try to improve things, and repeating until you find the sweet spot. You can imagine doing this,

FIGURE 2 Scatterplot of Example Data with Regression Line



but it sure doesn't sound like fun. In a fortunate turn, we don't have to do this graphically (where we measure things with a ruler); we can do it mathematically with the raw data. Even more good fortune, we can let computers do all the work for us (more on this later). As mentioned, the strength of a relationship between two variables is indicated on the scatterplot by how close the points are to a straight line. As we will see, in weaker relationships, the points are far from the line. The direction of the line (pointing up or pointing down) tells you direction of the relationship (positive or negative). If the line is completely flat, there is no relationship. Very important point: The apparent slope of the regression line (aside from the case where it is completely flat) does not indicate the strength of the relationship. It seems like it should, but it doesn't (aside from one special exception, with which we will not concern ourselves). As mentioned, the strength of a relationship between two variables is indicated on the scatterplot by the closeness of the points to a straight line, not the slope of this line. Why? The answer is that we can stretch or squash the *x*- and *y*-axes by a number of different methods to increase or decrease the slope of the line. The same data are displayed again in Figure 3, this time with different ranges on both axes.

FIGURE 3 Scatterplot of Sample Data with Adjusted Axes



That new slope sure looks amazing. But the strength of the association is unchanged. The correlation stays the same. So don't be fooled by the apparent slope of the regression line. Notice how I said that the *apparent slope* doesn't indicate the strength. If you were to compute slopes with the old slope = rise/run equation for the above graph and the previous one, you would find that the value is the same in both cases. By changing the range on the axes I've made the slope appear to be stronger. Always examine how close the points are to the line to assess the strength of the relationship. But this graphical stuff is just a visual representation of the data, something that we can eyeball to get a general idea of what is going on. To describe the strength of the association between two variables with any accuracy requires something more than a casual inspection of the raw scores or even a graph of these scores. We need a statistic to quantify the strength of the relationship. We have a few options. Before we discuss any of these statistics, let's discuss the properties a good measure of association statistic would have.

Measures of Association

What properties should a measure of association have? First off, it must accurately convey the desired information: strength and direction of the relationship. If it does't do that, then there's really no need to proceed with it. Furthermore, it should be sensitive to small differences in strength. One of the problems with evaluating strength with an examination of a scatterplot or dataset is that we are able to determine only the biggest of big picture ideas about strength ("It's kinda strong. Maybe, medium strong."). There is simply no way to be precise that way. A measure of association must be precise, or why bother?

Building on this, a good measure of association should be easy to interpret. That is, upon computing the coefficient, we should be able to determine, without any other information, whether the relationship is strong or weak, positive or negative. We should be able to see the number and instantly know what it means.

Finally, a good measure of association should have a design that makes some sort of logical sense. It should be more than just a magic box where the raw data is fed in the front, resulting in a coefficient falling out of the back end. I realize that this may not seem all that important, but it is.With these expectations set, let's examine our first measure of association, covariance.

Covariance

Covariance is not just our first measure of association, it's *the* first measure of association. Covariance is the start of it all. Aside from one incredibly annoying limitation, covariance is the simplest and most fundamental way to understand measures of association.

Covariance is like variance, but for a pair of variables. To understand covariance we must take a step backwards and discuss variance again. Variance quantifies differences among scores for a single variable (remember that if everyone has the same score, variance is zero). If you recall, variance (in the population form) is defined as the mean of the squared mean-deviation scores:

$$\sigma_X^2 = \frac{\sum (X - \mu_X)^2}{N}$$

Where:
μ_X is the mean of X.

As long as we're looking at the variance equation, I'll do a little algebraic manipulation and expand the squared part.

$$\sigma_X^2 = \frac{\sum (X - \mu_X)(X - \mu_X)}{N}$$

There, same equation, just presented a little differently. Just to refresh your memory a little more, a mean-deviation score is computed as the simple difference between a given score and the mean of that variable (i.e., $X - \mu_X$). A positive mean-deviation score indicates that the score is above the mean. A negative mean-deviation score indicates that the score below the mean. And a mean-deviation score of zero indicates that the score is, you guessed it, right at the mean.

To summarize how variance is computed, we transform every person's score into a mean-deviation score, square these mean-deviation scores, and compute the mean of these squared values.

Covariance is computed like variance – but for a pair of scores for each person. That is, instead of multiplying a each person's mean-deviation score on the X variable with itself (i.e., squaring), we multiply each person's mean-deviation score on X by his or her mean-deviation score on Y. To make this happen, all we need to do is make a small

modification to the variance equation listed above. Below is the equation for population covariance.

$$\sigma_{XY} = \frac{\sum (X - \mu_X)(Y - \mu_Y)}{N}$$

Where:
σ_{XY} is the population covariance of X and Y
μ_Y is the mean of Y.

Below is a dataset demonstrating the calculations for covariance.

Person	X	Y	(X-Mean$_X$)	(Y-Mean$_Y$)	(X-Mean$_X$)* (Y-Mean$_Y$)
John	12	1.1	-10.5	-1.45	15.225
Sal	23	2.8	0.5	0.25	0.125
Tim	24	2.9	1.5	0.35	0.525
Amy	31	3.4	8.5	0.85	7.225

The last column is the product of the mean-deviation scores. The mean of it yields the the covariance ($\sigma_{XY} = 5.775$).

Covariance Logic

How does the covariance equation work as a measure of association? Consider what we learned about bivariate associations: A strong positive association is obtained when people with high scores on one variable (e.g., X) have high scores on another variable (e.g., Y) and when people with low scores on X also have low scores on Y. When we say *high scores* and *low scores*, doesn't that sound like mean-deviation scores? (High or low compared to what? The other scores, with the mean being a great representation of the other scores.)

Now, all we need is a way to quantify the degree of consistency of these mean-deviation scores. Multiplying the two mean-deviation scores for each person results in a product which is only maximized when both scores are large numbers (either positive or negative). Take the mean of those products and you have a pretty sweet measure of association.

Let's talk a little more on how the mean of this product works. Consider that half of the the mean-deviation scores will be positive and half will be negative. If the people with high scores on *X* have high scores on *Y*, then you'll have two positive mean-deviation scores. Compute the product, and you have a nice, big, positive number (see Amy in previous dataset). Continuing with this example, if the people with the low scores on *X* have low scores on *Y*, then you'll have two negative mean-deviation scores. Take the product, and due to the old *negative times a negative equals a positive* property of numbers, and you'll have another nice, big, positive number (see John). The mean of all of the big, positive numbers is a big, positive number, indicating a strong, positive association. There's our index of strength and direction.

To continue to understand the logic of covariance, let's flip the previous scenario around. Now, the people with the high scores on *X* have the low scores on *Y* (and the converse). Thinking in terms of mean-deviation scores, that's a big, positive number multiplied by a big, negative number. Which results in a big, negative number. The mean of these is a big, negative number, indicating a strong, negative association.

And finally, what if there is no pattern? How does the covariance equation handle that? Some of the people with high scores on *X* have high scores on *Y*. Others have low scores on *Y*. Still thinking in terms of mean-deviation scores, that's some big, positive numbers multiplied by a big, positive numbers, resulting in big, positive products, *and* some other big, positive numbers multiplied by some big, negative numbers, resulting in big, negative products. Take the mean of these products, and you get a zero, indicating no association.

That's the logic of the covariance equation. That's how it quantifies the direction and strength of association. This statistic allows us to see how

these variables vary together, or co-vary (hence, the name covariance).

Totally irrelevant thought: What if the Y variable is just a copy of the X variable? Same scores and all. Doesn't that turn this part of the covariance equation: $(X - \mu_X)(Y - \mu_Y)$ into this: $(X - \mu_X)(X - \mu_X)$? A change which takes us back to the variance equation. As mentioned, covariance is like variance for a pair of variables.

Since there was a population and a sample form of the variance equation, you just know that there had to be a population and a sample form of the covariance equation. So here it is, the sample form of the covariance equation.

$$c_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Where:
c_{XY} is the covariance of X and Y in a sample of data.

There's really nothing else we need to say about this one. Other than the $N - 1$ thing and the use of the sample mean instead of the population mean, everything else is the same.

Now that we've described the mathematical basis for covariance, let's talk about what it does. As mentioned a number of times, covariance indicates the strength and direction of the relationship between two variables. A covariance of zero indicates no relationship between the variables. A positive covariance indicates a positive relationship between the variables, and a negative covariance indicates a negative relationship between the variables. The only problem with covariance as a measure of association is that it is difficult to understand just how strong or weak these relationships are. For one set of data a covariance of 41.4 might be weak, but for a different set of data a covariance of .83 might be very strong. It's all very annoying. This lack of a consistent standard for strong and weak relations is the major limitation

of covariance, and it is the principle reason why covariance is seldom used as an index of the association between two variables. (It does have other value in terms of summarizing data, but don't worry about that.) Good news though, there is a statistic that does indicate the strength and direction of the relationship between two variables in a standardized, easy to interpret fashion. And that statistic is the correlation coefficient.

Correlation

Correlation is, like covariance, a measure of association between two variables. Unlike covariance, correlation describes the association in a way that allows us to easily interpret the strength of the association. Correlation is, in essence, standardized covariance. Correlation is defined as the covariance divided by the standard deviations of each variable.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Where:
ρ_{XY} is the correlation of X and Y in a population.

Below is that sample version of the correlation equation.

$$r_{XY} = \frac{c_{XY}}{S_X \cdot S_Y}$$

Where:
r_{XY} is the correlation of X and Y in a sample of data. To reiterate, the symbol for sample correlation is r, variable name subscripts optional.

For both equations, the principle is the same: Correlation is covariance divided by the standard deviations of each variable. What purpose does that serve? Dividing by the standard deviations rescales the statistic so that the maximum and minimum values are always 1.0 and -1.0, respectively

(covariance maximums and minimums were a function of the product of the standard deviations; different standard deviations mean different maximums and minimums). Thus, a correlation of .6 always means the same thing in terms of strength, regardless of the standard deviations of the variables. That's the big advantage correlations have over covariance.

There are many types of correlation equations, but we'll focus on the most popular one, the Pearson Product Moment Correlation. If someone says that they correlated two variables, with no other information, they are talking about the Pearson one. If you use one of the weird ones (e.g., phi, tetrachoric, Spearman), you mention them by name.

The Pearson correlation summarizes the strength and direction of the association between two variables in a single number. The correlation coefficient ranges from -1 to +1. A positive coefficient means that the relationship is, you guessed it, positive. A negative coefficient indicates a negative relationship. A -1.0 correlation indicates a perfect negative relationship and a +1.0 correlation indicates a perfect positive relationship. A 0.0 correlation indicates no relationship between the two variables. Thus, the strength of the relationship is indicated by how close the number is to +1 OR -1. A correlation of -.8 is just as strong as a +.8. I hope that it is clear that the sign of the correlation is irrelevant to the strength of association. The direction of the relationship is useful information worth knowing; it is just different information than the strength of the relationship.

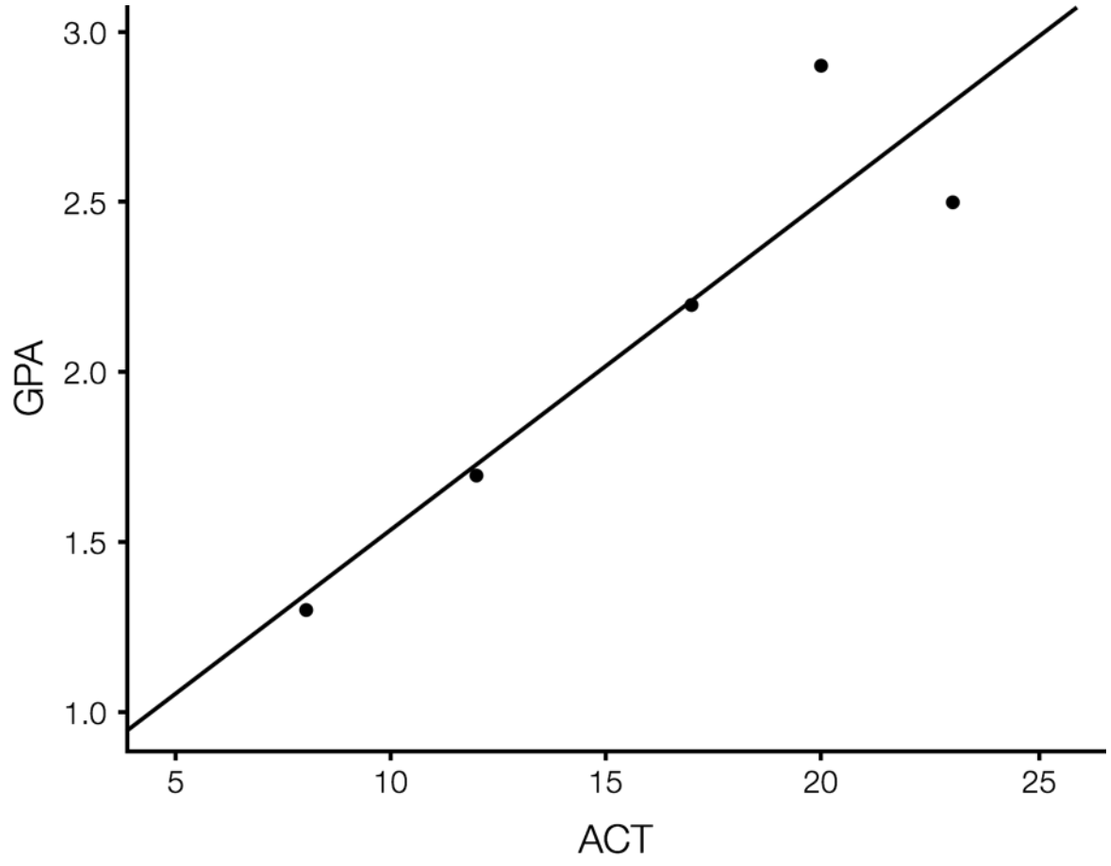
Consider the following dataset.

Person	X (ACT)	Y (GPA)
Frank	8	1.3
Kevin	12	1.7
Gianni	17	2.2
Warren	20	2.9
Judy	23	2.5

As you can see in Figure 4, the rank order, although good, is not perfectly consistent. The person with the highest score on *X*, Judy, has the second highest score on *Y*. The person with the second highest score on *X*, Warren, has the highest score on *Y*. They are out of order. Everyone else falls in line (third on *X* is third on *Y*, fourth is fourth, etc.).

Clearly the trend is positive, but compare this graph to any of the scatterplots of the previous dataset (Figure 2). Notice how the points in our new

FIGURE 4 New Dataset, Weaker Association



scatterplot are not as close to a straight line. Weaker association. Computing the correlation confirms what we already know, $r = .92$. Still very strong, but weaker than the previous dataset.

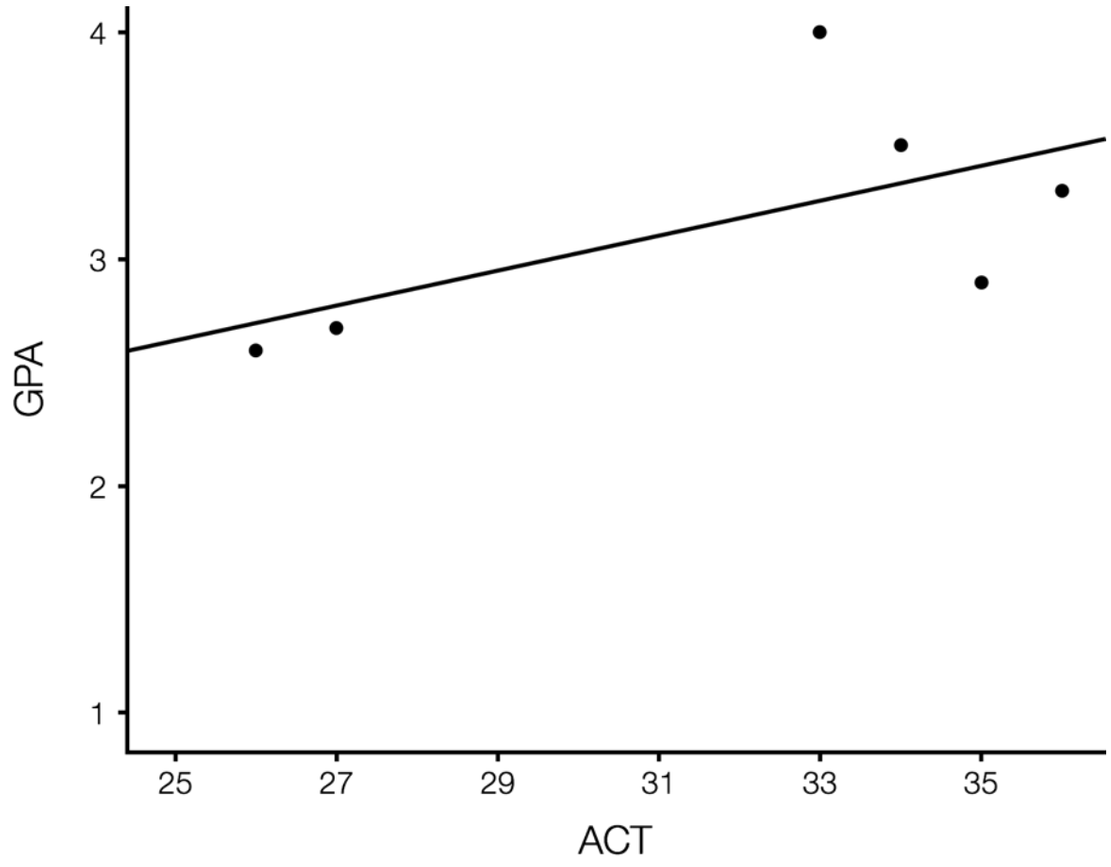
Time for a new example.

Person	X (ACT)	Y (GPA)
Rusty	26	2.6
Buck	27	2.7
Jeff	33	4.0
Dale	34	3.5
John	35	2.9
Margaret	36	3.3

Now we see even more exceptions to perfect rank ordering (Figure 5). The person with the highest score on *X*, Margaret, has the third highest score on *Y*. The person with the second highest score on *X*, John, has the fourth highest score on *Y*. More exceptions abound, but in spite of them, we can still see a general trend: Higher scores on *X* are associated with higher scores on *Y*.

The line points up, but the points are even further from the line than we have seen before. Thus,

FIGURE 5 Even Newer Dataset, Even Weaker Association



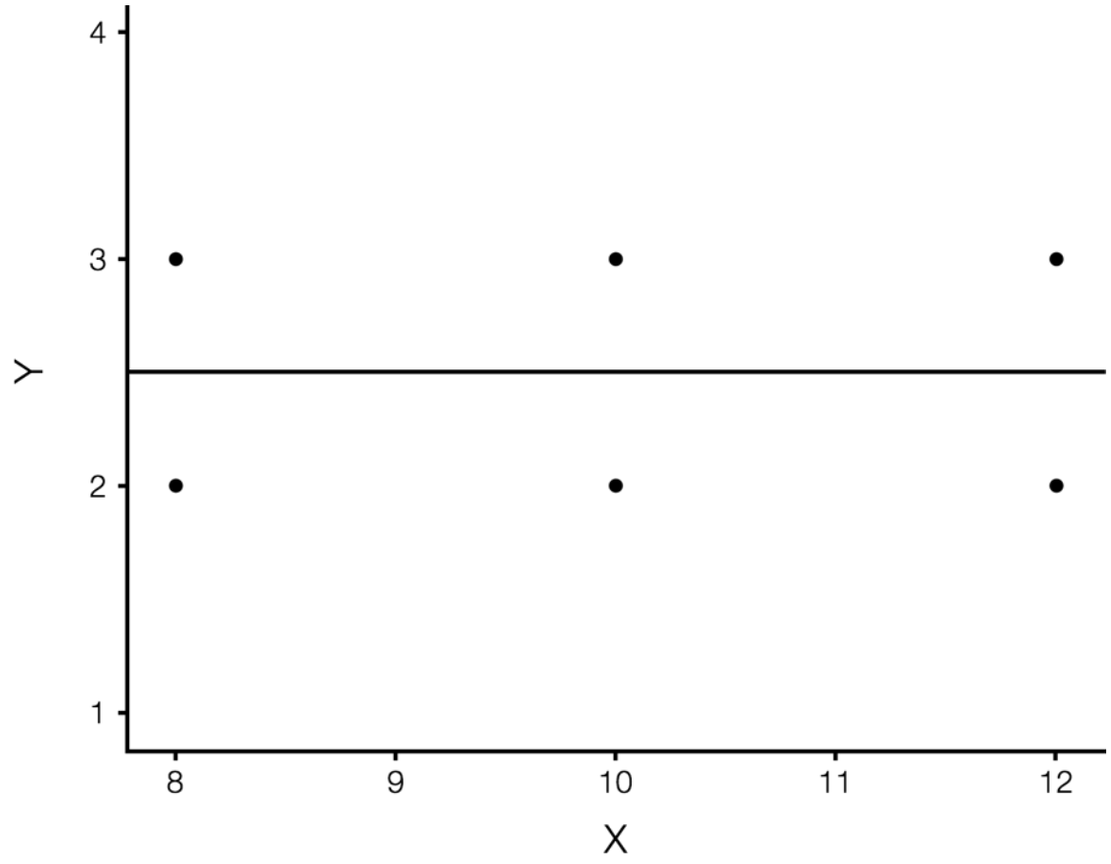
we have a positive association that's not perfect. How strong is it? $r = .61$. So it's positive and strong, but weaker still than the previous datasets.

You might be wondering what a zero correlation looks like. Well, wonder no more.

Person	X	Y
Hunter	8	3
Lonny	8	2
Charles	10	3
Craig	10	2
Danny	12	3
Kendall	12	2

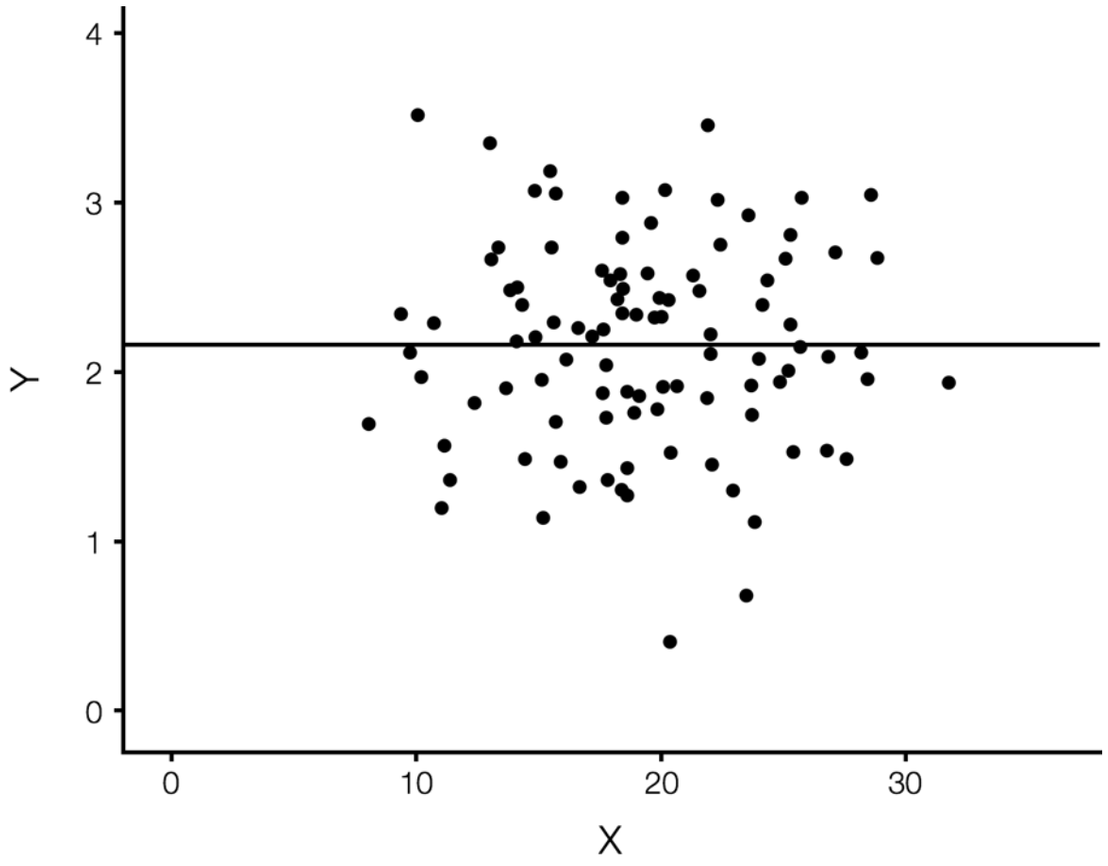
What do we see? No clear trend. High scores on *X* are associated with both high and low scores on *Y*. Low scores on *X* are associated with both high and low scores on *Y*. The scatterplot is shown in Figure 6 and looks like a rectangle of scores. I'll bet you didn't know what a rectangle of scores looked like before now. I'll also bet that you didn't care to know. And you still don't.

FIGURE 6 Zero Correlation Scatterplot



Of course, a six person dataset makes for a fairly uninteresting scatterplot when the correlation is zero, but, hopefully, the point is made: There is no trend in the data.

FIGURE 7 A Zero Correlation Scatterplot from a Sample of Conventional Size



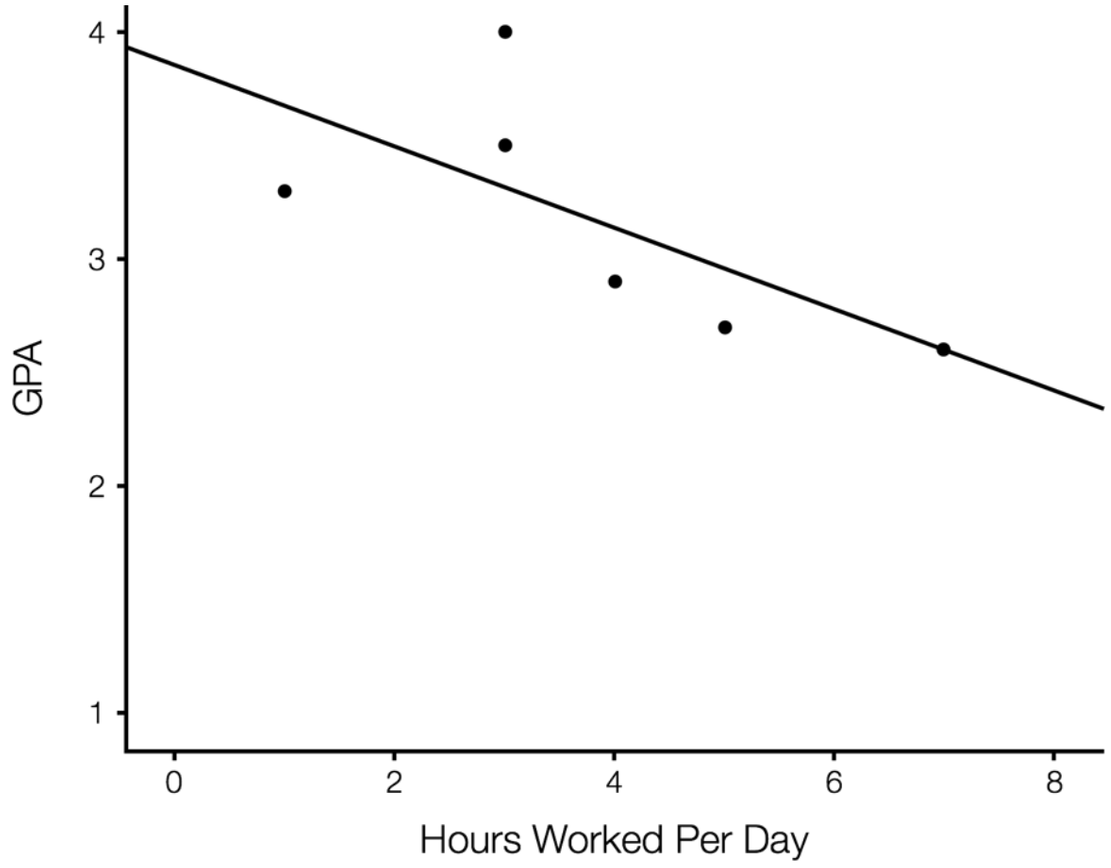
When you have a larger dataset, a zero correlation scatterplot resembles a circle (Figure 7). Notice how the regression line is perfectly flat. No slope at all. This is the land of $r = 0.0$. A bleak and desolate land. Unfit for both man and animal. No association of *X* and *Y* of any kind.

Finally, how about a negative correlation.

Person	X (Hours Worked)	Y (GPA)
Nick	7	2.6
Al	5	2.7
Evan	3	4.0
Eddie	3	3.5
Ernie	4	2.9
Kelly	1	3.3

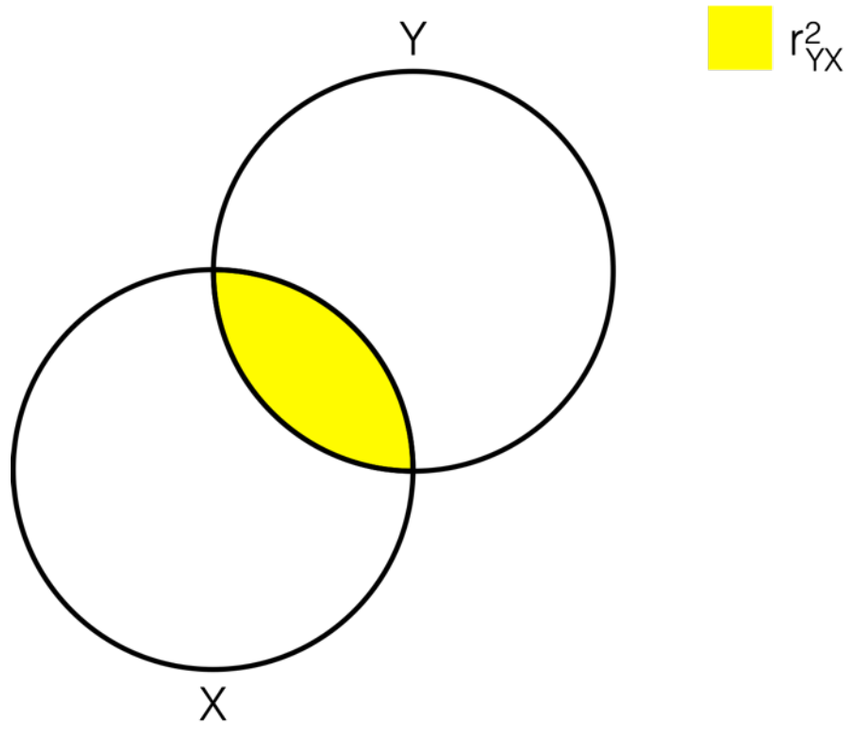
We can see a clear (although not perfect) trend: Higher scores on *X* are associated with lower scores on *Y*. The scatterplot is shown in Figure 8 and is different from our previous examples.

FIGURE 8 Negative Correlation Scatterplot



The line of best fit points down, indicating a negative association; the points are fairly close to the regression line, indicating a strong association. So we see a strong, but not perfect, negative association. The actual correlation is r = -.68.

FIGURE 9 Venn Diagram Illustrating Relationship between X and Y



At this point, it's time to introduce another way to illustrate the association between variables. If you like Venn diagrams, and who doesn't, one is shown in Figure 9. Venn diagrams illustrate the relationship between various concepts. When applied to correlations, the circles represent the variance of each variable. The overlap of the circles indicates the degree of association. Greater over-

lap indicates greater associations. To get technical, the percent of the area of Y overlapped by X represents the squared correlation between the two variables (i.e., r^2_{XY}). We'll talk more on squared correlations later, but it's not like there's a lot of mystery. You have a correlation. You square it. You get r^2.

Computing Correlation Coefficients

As mentioned, there are a number of types of correlations, but we'll stick with the ever popular Pearson correlation. Since we live in a computer age, there is little to be gained by focusing on equations. Little, but not nothing. There are a few different, but equivalent, versions of the equation for the Pearson correlation. We've already seen one that starts with covariance. Let's examine the most intuitive form of the Pearson correlation equation, the average product of z scores. Listed below is the population version of it.

$$\rho_{XY} = \frac{\sum (z_X \cdot z_Y)}{N}$$

It's fairly simple – just compute the product of the z scores for each person, compute the mean of those products, and you're done.

To refresh our memory on z scores, the population form of the z score equation is listed below.

$$z_X = \frac{(X - \mu_X)}{\sigma_X}$$

And, of course, if we want to compute z scores for Y, the equation is be the similar, only with Y substituted for X at every opportunity.

Just for fun, let's take these z score equations for X and Y and substitute them into the correlation equation from above. Here's what we obtain with those substitutions:

$$\rho_{XY} = \frac{\sum (\frac{(X - \mu_X)}{\sigma_X} \cdot \frac{(Y - \mu_Y)}{\sigma_Y})}{N}$$

It's still the correlation equation, but it looks familiar. Where have we seen something similar? That's right, take away the standard deviations (σ_X,σ_Y), and it's the equation for population covariance.

$$\sigma_{XY} = \frac{\sum (X - \mu_X)(Y - \mu_Y)}{N}$$

Do you see it in the above two equations? The only differences between the correlation (ρ_{XY}) and covariance (σ_{XY}) equations are the standard deviations in the former.

Here's a thought exercise: What if X and Y were standardized? The standard deviations would both be 1.0. And since anything divided by 1 equals itself, the standard deviation parts of the correlation equation disappear, leaving us with what we see in the covariance equation. (One quick lesson from this is that for standardized data, covariance equals correlation.)

As mentioned earlier, covariance is computed as the average of the products of mean-deviation (i.e., $X - \mu_X$) scores for each person. Correlation is computed as the average of the products of the z scores for each person. And what's the difference between a z score and a mean-deviation score? A division by a standard deviation. I hope that it's clear to you that correlation and covariance are very similar statistics. But correlation is better. There, I said it.

Those were population versions of the equations. For reasons that should be obvious by now, it will be much more useful if we discuss the correlation equation designed for samples. And here it is.

$$r_{XY} = \frac{\sum (z_X \cdot z_Y)}{N - 1}$$

What's the difference? Well, there's the symbol for correlation. It's now r instead of ρ. So, there's that. The only other difference is the denominator.

It's $N - 1$ instead of just N. Does this look familiar? It should. This is the variance story all over again. When computing the variance of a population of data, the denominator is N. When computing the variance of a sample of data, the denominator is $N - 1$. It's the same pattern with the Pearson correlation equation: the N denominator for populations, the $N - 1$ denominator for samples. So, when computing z scores for the sample correlation equation (sample z score equation repeated below), be sure to use the appropriate $N - 1$ variance equation. With samples, it's $N - 1$ denominators all the way down.

$$z_X = \frac{(X - \bar{X})}{S_X}$$

Of course, we let computers do the dirty work for us, and they use sample versions of equations for everything. But just in case you have to do computations by hand, you have enough information to do it right.

How Correlations Work

Let's put this all together so that we can really understand what makes correlations tick. Correlations are measures of association between two variables. When people who have high scores on one variable also have high scores on the other variable (and vice-versa), you get a strong, positive correlation. The Pearson correlation equation quantifies the relationship by computing the mean cross-product of z scores. Interactive 1 demonstrates this process.

INTERACTIVE 1 How Correlations Work

Correlation and Causation

Correlation does not equal causation. It's important to remember that a correlation is just a statistic that describes the association between two variables. *Why* these variables are associated is another matter. *Why* is an issue of causality. In general, our statistics can't address causality. It is our research design that allows us to address causal issues. As but one example, consider the ACT/College GPA correlation. There is a positive correlation of about .5 between these two variables. Does that mean that your performance on the ACT causes your college performance (*X* causes *Y*)? Probably not. Does that mean that your college performance causes your ACT performance (*Y* causes *X*)? We can safely rule this out based on logic: ACT performance is measured months before college performance even begins. Statistically, the *Y causes X* inference is as valid as the *X causes Y* inference. It is our research design that allows us to rule out *Y* causing *X* in this case. So we've covered the causality issue in both directions. There is, however, a third possibility. It is possible that a third variable, which we'll call *Z* , is causing performance on both *X* and *Y* – making *Z* responsible for the correlation between *X* and *Y*. What is this third variable in our ACT-GPA example? Let's pick one: study habits. People with good study habits do well on the ACT and do well in college. People with poor study habits generally do poorly on both. So, it appears that *Z* is responsible for the correlation. No guarantees, but if I was betting person, which I am not, I'd bet on *Z*. (Just to be complete, there is also a fourth option in which *X* causes *Z* which causes *Y*, making *X* an indirect cause of *Y*. Don't worry about it, though. The previous explanation is far more relevant.)

New example: Ice cream sales (*X*) are correlated .7 with shark attacks (*Y*) at a certain seaside resort. Which seaside resort? That information is classified. Is *X* causing *Y*? Maybe, if people are eating a bucket of ice cream and then going swim-

ming right away. Is it possible that the sharks are attracted to the smell of ice cream? Can they even smell it? Does the flavor matter? All good questions, but let's switch gears. Is *Y* causing *X*? That is, are the shark attacks causing people to buy ice cream? Maybe the survivors of the shark attacks like to celebrate cheating death with some mint chocolate chip. Statistically, both are equally valid explanations – do you see why research design is so important? Also, do you see the dangers of a blind application of statistics (i.e., devoid of logic)? Now is it possible there is some third variable at work here? Yeah, probably.

As you know, a correlation of zero indicates that there is no relationship between *X* and *Y*. And you know that +1 and -1 indicate perfect relationships. But what are industry standards for strong, medium, and weak correlations? The classic resource on this issue is Cohen (1992). Cohen's standards for correlational strength are as follows: small is .10, medium is .30, and large is .50. Naturally, the same rules apply to negative correlations. As Cohen stated, .10 is small; it's far too weak to be useful under most circumstances. So consider .30 to be the minimum decent value for a correlation.

Significance Testing in Correlation

At the beginning of this book, I mentioned that we wouldn't concern ourselves with significance testing. And we won't. At least, not much. Computers do them for us, but it helps to see the equation. Before I explain how to conduct a significance test for a correlation coefficient, I need to explain the general concept of significance testing. First, we need to review three terms from earlier: sample, population, and sampling error. I'm sure you recall that we measure samples because it's inconvenient or impossible to measure an entire population. We analyze data in our sample and make inferences from the sample to the population (e.g., 55% of the people in our sample watch

football on TV; thus, we estimate that 55% of the people in the population watch football on TV). Unfortunately, measuring a sample instead of the entire population leads to problems. The results that we find in our samples will not be a perfect match to the population statistics. The difference between the two is called sampling error, and it is the price we pay for laziness.

Given our knowledge of sampling error, it should be clear that we should not infer too much from our samples. Quick example: let's say we collect a sample ($N = 127$) of ACT scores from high school athletes. We analyze the data and find that soccer players have a mean score of 20.5, and tennis players have a mean score of 20.3 (yes, it's a pointless study). It is obvious that soccer players outperformed tennis players in our sample of 127 students. But should we make an inference to the population and say that soccer players score higher than tennis players on the ACT? Probably not, you say, since the difference between the mean scores is so small. Good call. It is a mistake to think that a small difference in our sample statistics indicates that there is a difference between the groups in the population. The small difference in our sample could be due to sampling error. Now what if there was a big difference in the sample means (let's say that the soccer players outscored the tennis players by 11.5 points)? Does this large difference in sample scores allow us to conclude that soccer players outscore the tennis players in the population? Yes, it is likely that they do (assuming certain other things with which we will not concerns ourselves at the moment).

Where do significance tests fit in? Significance tests (also called *inferential* statistics) are used to analyze the sample characteristics and indicate when it is wise (or unwise) to make inferences to the population. They tell us if we can conclude that a certain characteristic (e.g., a difference between test scores of girls and boys) actually exists in the population. Significance tests are probabil-

ity analyses, and give an answer like, "There is only a three percent chance that a result like the one we found in our sample could have been found if there truly was no difference in the population. Therefore, we conclude that there is a difference in the population." (It helps if you read that sentence with a deep, authoritative voice.)

So what about significance tests for correlations? It's the same story except now that we examine the correlation in our sample (r_{XY}) and use it make inferences about the relevant population correlation (ρ_{XY}). Example: let's say we collect a sample of college students ($N = 93$) and find that time spent playing video games is positively correlated with GPA, $r = .07$. Sure, it's a weak correlation, but it's a positive correlation. It is indisputable that in our sample people who spent more time playing video games had a higher GPA. Go ahead, try and dispute it. Can we then conclude that in the population more time spent playing video games is associated with higher GPAs? I hope you're shouting, "No, our sample correlation is likely influenced by sampling error! The population correlation could be zero for all we know! The sample correlation is only slightly greater than zero!" That's enough shouting for now. Your instincts are correct. We need to conduct a significance test to determine whether our sample correlation is large enough to allow us to conclude that the population correlation is not zero.

Now that we have the general significance testing issues out of the way, how does the significance test of a correlation work? It's a *t* test, and all you need to know to perform the test are the sample correlation and sample size.

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{N - 2}}}$$

Pretty simple. Once you have obtained your *t* value from the sample statistics (we often call this the *obtained t*), use a *t* table to find the critical *t*

(with $N - 2$ degrees of freedom) and compare the two. If the obtained t is greater than the critical t, then we say the sample correlation is significant (note: the previous term is statistician slang) and conclude that the hypothesis is supported. If our hypothesis specified a positive direction, then we would conclude that the population correlation is greater than zero.

Regression Introduction

First off, we need to address this word *regression*. Regression analysis is only vaguely related to the regular words *regression* or *regress*. The name is not important. They could have named it with a nonsense word like shniffle. Or omegnacruz. Quavelcon analysis sounds pretty cool. Names aside, regression is a statistical procedure that describes the relationship between two variables (like correlation) and allows us to use this relationship to predict a person's score on Y given their score on X (unlike correlation). To reiterate: Correlation is a measure of association between two variables, whereas regression provides measures of association *and* a method for predicting scores on one variable given scores on another. With regression, you get the bonus plan.

A few miscellaneous issues. First, regression and correlation are so closely related that it's hard to tell which one is derived from which (i.e., which came first?). I like to conceptualize it as regression is an extension of correlation – it starts with correlation's function and takes it to a new level. That's how I like to think of it, but it doesn't matter a whit. Next, in this chapter, we won't address the *sample* versus *population* form of a statistic anymore. It was fun and all, but we're done with that. We'll just assume everything is a sample and present our equations accordingly. And you know enough by now that, if you had to generate a population version of any of these equations, you could figure it out without breaking a sweat. Finally, a note about terminology. We'll start with

a discussion about something that is often called *simple regression*. The full, proper name for what is described in the next few sections is *bivariate linear regression*. Simple regression is a shorter name, and it fits well because bivariate linear regression is the simplest form of regression. Bivariate means two variables (X and Y). Linear means linear (as in, the regression line is a straight line). And regression means, well, nothing helpful, as mentioned a few paragraphs ago.

Regression Basics

To get the basics down, consider the equation in simple, ordinary least squares regression in which Y is the dependent variable and X is the independent variable:

$$Y' = bX + a$$

Where:
b is the regression coefficient; the slope of the regression line; the weight applied to scores on X to get the best possible prediction of Y.
a is the y-intercept; a constant needed for scaling purposes.
Y' is the predicted value of Y; note that this isn't Y (a person's actual score on Y), but rather our prediction for Y for that person.
X is the person's score on X; substitute a person's score on X to get their predicted value for Y.

An example should help. Let's say that we know that $b = 1.3$ and $a = 8.2$. Don't worry about how we obtained these numbers. In this case, $Y' = 1.3X + 8.2$ (or you could say $Y' = 8.2 + 1.3X$). For the following dataset, we can compute Y' for each person by inserting each person's X score into the above equation.

Person	X	Y’
Hal	5	14.7
Fred	2	10.8
Eddie	6	16.0
Joe	8	18.6
Charles	2	10.8

That’s it in all its glory. Predicting a person’s *Y* score with a regression equation is a simple algebraic exercise (this is referred to *actuarial* or *statistical prediction*; in contrast there is *clinical prediction* which is a judgmental method). Note that the same *X* will always result in the same predicted Y (see Fred and Charles). Our opinions don’t count. Just the equation and the data.

Now let’s say that somehow we know each person’s actual score on *Y*. We can compare these actual *Y* scores to our predicted *Y* scores.

Person	X	Y’	Y
Hal	5	14.7	16
Fred	2	10.8	9
Eddie	6	16.0	10
Joe	8	18.6	22
Charles	2	10.8	14

As we can see, our predictions were pretty close for some of the people (Hal and Fred) and were way off for the others (Eddie, who did a lot worse than we predicted, and Joe and Charles, who did a lot better than we predicted). The difference between the actual *Y* and the predicted *Y* is called the residual, and it shows the amount of error in the prediction of *Y* for each person. Also note that Fred and Charles had the same predicted *Y* but had very different scores on actual *Y*. This Fred/Charles situation illustrates how the same scores on *X* will always result in the same predicted *Y*; however, their actual scores on *Y* will likely turn

out to be different (unless our prediction is perfectly accurate).

Person	X	Y'	Y	(Y – Y')
Hal	5	14.7	16	1.3
Fred	2	10.8	9	-1.8
Eddie	6	16.0	10	-6.0
Joe	8	18.6	22	3.4
Charles	2	10.8	14	3.2

I hope that it is obvious that we like it best when there are no errors of prediction. Such a situation would indicate that our predictions were perfectly accurate. But that doesn't happen in real life.

Let's talk about the accuracy of predictions made with regression analysis. Anyone can make predictions. We want to make predictions when we have a good chance of being accurate. How can we know whether these predictions will be accurate? That's where correlation enters the picture.

Stronger correlations between *X* and *Y* lead to more accurate predictions. (In the above dataset, $r = .65$.) A perfect correlation (+1 or -1) would give us perfectly accurate predictions (0.0 residuals for all people). Of course, perfect correlations don't happen in the real world, but you get the idea.

Now seems like a good time to learn whence *b* and *a* are derived (i.e., where they come from).

$$b = r_{XY}(\frac{S_Y}{S_X})$$

$$a = \bar{Y} - b\bar{X}$$

That's it. You only need five very basic statistics to generate a regression equation and start predicting *Y*: the means of *X* and *Y*, the standard deviations of *X* and *Y*, and the correlation between *X* and *Y*. Fairly simple.

A Regression Thought Experiment

An examination of the equation for b reveals something interesting: The most important part of b is the correlation between X and Y. The standard deviations are just scaling terms. (Consider that if the data are transformed to z scores, then S_X and S_Y both equal 1.0 and are irrelevant.)

Here's an interesting thought experiment: What happens if $r_{XY} = 0$? In this scenario, we won't even have to standardize our data. It can be raw data. Using the above equations, when $r_{XY} = 0$, $b = 0$ (because anything multiplied by zero is zero). And if $b = 0$, $a = \bar{Y} - 0\bar{X}$, which simplifies to $a = \bar{Y}$. And, thus, the regression equation is $Y' = 0X + \bar{Y}$, which simplifies nicely to $Y' = \bar{Y}$. To restate: if $r_{XY} = 0$, then the regression equation is $Y' = \bar{Y}$. Thus, predicted Y is the mean of Y for all scores on X. It doesn't matter what your score on X is, we predict the mean of Y for you. Why? Because there is no association between X and Y. Thus, why should I care about your X score in my prediction of Y? That zero correlation tells us X is irrelevant. Earlier, we said that b tells us how much to weigh scores on X to get the best possible prediction of Y. If $r_{XY} = 0$, then X doesn't matter, and I should give it no weight in my prediction of Y (hence, $b = 0$ in this scenario).

To the converse, what if $r_{XY} = 1.0$? To make matters easy on ourselves, let's also make X and Y standardized data so that the means are 0.0 and the standard deviations are 1.0. Inserting these numbers into our equations for b and a yields $b = 1.0$ and $a = 0$. (Try it, you'll see.) Thus, our regression equation is $Y' = 1X + 0$, which simplifies to $Y' = X$. With this equation, if X is 2.3, then Y' is 2.3. And if X is -1.9, then Y' is -1.9. Predicted Y is exactly as high or low as X.

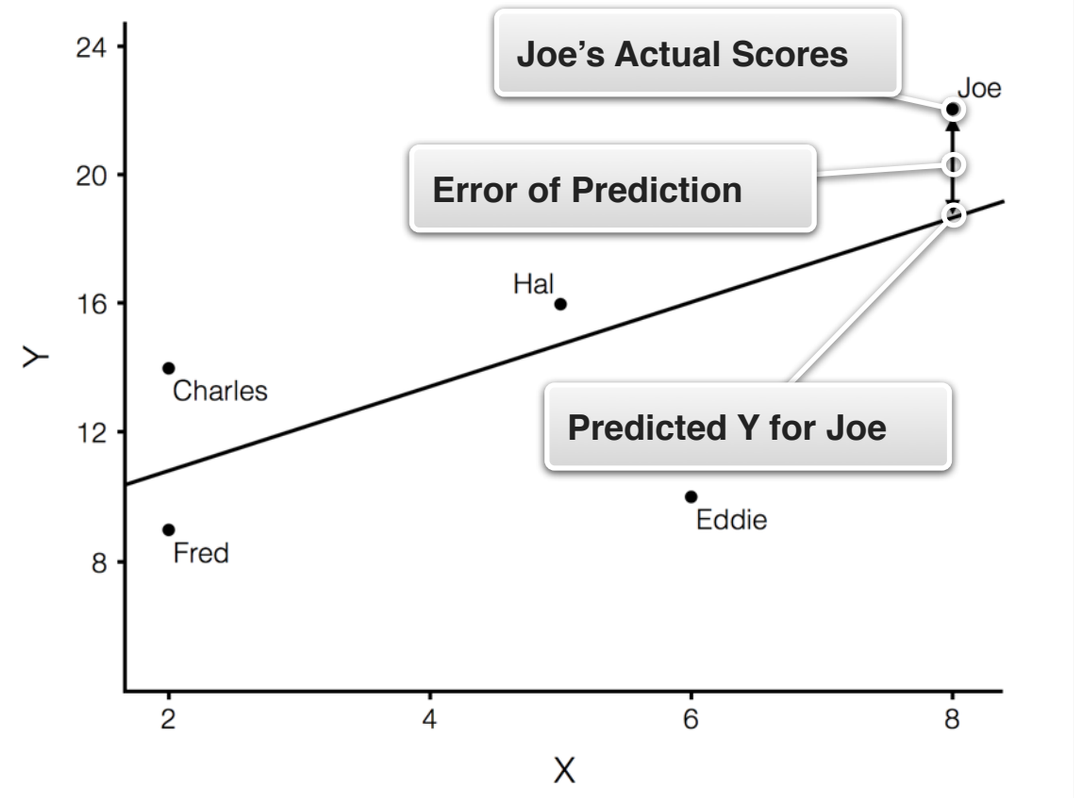
Conclusions from our little thought exercises: The relationship between X and Y strongly determines the types of predicted Y scores generated

from a regression equation. If $r_{XY} = 0$, then all predicted Y values are the same, regardless of the score on X. If r_{XY} is weak, then all predicted Y values are close to mean, even for people with extremely high or low scores on X. If r_{XY} is strong, then people with extremely high or low scores on X will have very high or low predicted Y values.

The Regression Line

Remember that line of best fit in the correlation graph? It was also called the regression line. It graphically represents Y' values for all scores on X. You can draw the line by plugging all possible values of X into the regression equation, obtaining Y' for each X, and graphing each of the X, Y' points. Or you could just draw the line using the regression coefficient b as the slope and the a as the y-intercept. Our most recent dataset is graphed in Figure 10, and it includes the regression line. Note that the error of prediction (or residual) is indicated graphically by the vertical distance between each point and the regression line. Bigger distances mean worse prediction (more error). Let's examine Joe's data. Joe has a score of 8 on X. His predicted Y is 18.6. His actual Y is 22. Thus, his error of prediction is +3.4. That is, he performed 3.4 points better than we predicted. By

FIGURE 10 Actual Y, Predicted Y, and Errors of Prediction (e)



comparison, Hal has a much smaller error of prediction (1.3). Someone with scores located right on the regression line would have an error of prediction of zero. The old rule from correlation-land is relevant again: Stronger associations (which lead to more accurate predictions) are those with points closer to a straight line. Correlation and regression, two sides of the same coin. Or maybe they are the same side of the same coin.

A quick note on interpreting the regression coefficient. The regression coefficient, *b*, tells us something useful, and unique, about the relationship between *X* and *Y*. For simple regression, *b* indicates the expected change in *Y* given a one point change in *X*. Here's an example. Let's say we conduct an experiment where we assign people to varying levels of study time and then measure their test performance. We regress test scores (*Y*) on study time (*X*, measured in hours) and obtain the regression equation $Y' = 22X + 11.5$. In this equation, *b* is 22. Using our interpretive rule, for every one hour increase in study time, we expect to see a 22 point increase in test scores. We may find this information to be very useful in evaluating the relationship between study time and test performance. Also, the effects of *b* are cumulative; a three hour increase in study time will be expected to lead to a 66 point increase in test performance. Nice.

So the regression coefficient can be a useful indicator of the strength of the association between two variables. Just like correlations. Sometimes a regression coefficient can be every bit as useful (arguably more so) than a correlation. When is that? It all depends on whether the dependent variable is expressed in a meaningful metric. What kind of metrics of measurement are meaningful? Anything that has real world relevance, such as time (e.g., to complete a task), number of something (e.g., mistakes), dollar value (e.g., of items sold), and so on (e.g., and such and forth). Let's say that the dependent variable is expressed in dollars. If

the regression coefficient is 50, then for every one point change in *X*, we expect *Y* to increase by $50. Based on the nature of the experiment, it will be easy to interpret whether $50 is a meaningful change (and thus, a strong relationship) or a trivial one.

As long as we're discussing strength of association, remember how we mentioned in our discussion of correlation that we shouldn't use the slope of the line on the scatterplot as an indicator of the strength of the association? Mostly because the *x*- and *y*-axes can be easily manipulated to produce the appearance of a strong slope. Remember that? Well, here's the thing. Even though it's unwise to use apparent slope of the regression line on the scatterplot as an indicator of the strength of association, it is fine to use the regression coefficient as an indicator of the strength of association. How is that OK, you're thinking? I once had a perceptive student ask me this very question. Well, we know why a visual inspection of the slope of the line on the scatterplot is a bad idea. But wait, didn't we just learn that the regression coefficient is the slope of the regression line? How then is the regression coefficient useful? The answer is that it's not as easy to manipulate the regression coefficient. You can't just change the scale on the axis of some graph and get the desired effect. You would have to change the scale of the data itself (e.g., multiply all of the scores on the dependent variable by 10). Such a change would be obvious. In short, if one wants to produce the appearance of a strong relationship, it's more difficult to manipulate the regression coefficient than it is to manipulate the apparent slope of the regression line on the scatterplot. The former is a number. The latter is a visual representation of that number.

To show how a regression has a real-world usefulness, let's pretend that we are in charge of admissions of a certain college. We'll call it Enormous State University (or ESU). At ESU we are considering using the ACT for freshman admis-

sions. Somebody somewhere (maybe at arch-rival Enormous Tech) did a study and found a .5 correlation between ACT scores and college GPA. Based on that correlation, we decide to use the ACT at ESU. Naturally, we also need the means and standard deviations of X and Y to set up our regression equation. Once we get these data, we obtain a regression equation of $Y' = .1X + .5$. For every high school student that applies, we plug his or her ACT score into our equation, which generates a predicted Y for each person. If the predicted Y is high enough (say, greater than 2.0), then we admit the student. If not, then we send him the other letter.

Significance Testing in Regression

Significance testing just won't go away. Significance testing in regression is very similar to significance testing in correlation, with a twist. First off, the r_{XY} we obtain from a regression analysis is the same as the r_{XY} we obtain from a correlation analysis (except that it can't be negative). That much should be clear by now. Technically speaking, regression gives us r^2_{XY} (yes, it's just the correlation squared), but r_{XY} is just a square root button on the calculator away. As these correlations are the same magnitude, it should not be a surprise that the outcomes of the significance tests are the same. Again, with a twist.

In regression, the significance test we conduct is actually a test of r^2. The equation is as follows.

$$F = \frac{r^2_{XY}/k}{(1 - r^2_{XY})/(N - k - 1)}$$

Where:
r^2_{XY} is the squared correlation between X and Y.
k is the number of independent variables.
N is the sample size.

Because this is simple regression, there is only one independent variable, meaning $k = 1$. This test is an F test with k, $N - k - 1$ degrees of freedom.

Here's the cool part: The F test of r^2 is identical to the standard t test of r as long as the t test is a two-tailed test. Don't forget that last part: These two significance tests yield the same result if the t test is a two-tailed test. (Note: There is no tailedness to an F test. A clue to this can be found by noticing that, quite obviously, r^2 cannot be negative – a positive relationship between X and Y and a negative relationship between X and Y of the same magnitude will result in the same r^2.)

So that was pretty easy. There are other significance tests in regression analysis, but we won't worry about them. For a more thorough treatment of this topic, consult my book, *Fundamentals of Correlation and Regression*.

Understanding Strength of Association in Regression

In this chapter, we've mentioned three ways to assess the strength of association in regression analysis: r, r^2, and b. All three have their own significance tests. But strength of association isn't significance testing. In using these statistics to understand strength of association, all three statistics have their merits. All have weaknesses in this area as well.

As mentioned earlier in the chapter, Cohen's (1992) standards provide useful guidelines for assessing the strength of a regular (i.e., un-squared) correlation. We also have discussed the use of the regression coefficient as an index of the strength of association (earlier in this chapter). The regression coefficient is very useful, arguably more useful than a correlation, if the dependent variable is in a meaningful metric, such as time, money, number of accidents. As discussed earlier, the regres-

sion coefficient can be interpreted as follows: For every one point change in scores on *X*, we expect score on *Y* to change by *b* points. For example, the equation $Y' = 35X + 130$ tells us that for every one point change on *X*, we expect scores on *Y* to increase by 35 points. If *Y*, the dependent variable, refers to days spent working, then this *b* of 35 has real meaning. We expect someone who scores a 10 on the test to work 35 days longer than someone with a score of 9.

Finally, let's discuss the use of r^2 as an index of the strength of association. No doubt about it, r^2 has a cool name: coefficient of determination. And it has an impressive definition: r^2 indicates the percent of variance in *Y* explained or accounted for by *X*. So if *X* and *Y* are correlated .5, then r^2 is .25. This r^2 means that 25% of the variance in *Y* is explained by scores on *X*. Fans of simple math will note that this also means that 75% of the variance is not explained by *X*. Our .5 correlation (which Cohen describes as "strong") doesn't sound so strong anymore. Them's the breaks when you square numbers between 0 and 1. They get smaller.

It appears that we have a conundrum. A .5 correlation is strong, so says Cohen. But a .5 correlation fails to explain 75% of the variance in *Y*, so says r^2. A relationship can't be strong and weak at the same time. What's going on? The answer is hiding in plain sight. The definition of r^2 states that r^2 indicates "the percent of variance in *Y* accounted for by *X*." The word *variance* is where the problem occurs. As we learned in some previous chapter, variance is in squared units (e.g., squared ACT points). Thus, if ACT scores are correlated .5 with GPA, the r^2 method of assessing relationship strength is saying that "differences among squared ACT points explain 25% of the differences in squared GPA points." This is not at all helpful. What we want is a way to understand how two variables are related to each other while retaining the regular metric of measurement (i.e., un-

squared points). Brogden (1946) demonstrated that the regular, un-squared correlation is linearly related to how well one variable predicts another. Using our example, a variable correlated .5 with *Y* predicts *Y* half as well as a variable perfectly correlated with *Y*. A variable with a .4 correlation is 40% as efficient at predicting *Y* as is a variable with a 1.0 correlation. And so on.

Before I take any criticism from the gallery for my disdain for r^2 as an index of the strength of association between *X* and *Y*, let me say the following. I understand that predictive efficiency isn't relevant to every discussion regarding strength of association. However, when predictive efficiency is relevant, "percent of variance accounted for" is wholly inappropriate for understanding the strength of association. That said, even for regression analyses not focused on prediction (i.e., causal research), interpreting *r*, instead of r^2, is still the better way to understand strength of relationship. A correlation of .5 is 50% as strong as a perfect relationship. An interpretation of r^2 would lead you to believe that it is only 25% as strong because *X* only accounts for 25% of the variance *Y*.

Conclusion: r^2 may have a cool name and a definition that sounds useful, but it is not the best way to understand how well two variables are related to each other. Stick to statistics that remain in the original metric of measurement: *r* and *b*.

Regression Analysis Summary

Regression analysis extends the concept of correlation and applies it in new ways. A correlation coefficient simply describes the relationship between two variables. Like correlation, regression analysis describes the relationship between two variables. This description can be done with any of three different measures of association: *r*, r^2, and *b*. Unlike correlation, regression analysis can be used to predict scores on the dependent variable based on scores on the independent variable.

These predictions are made based on the association between *X* and *Y* (and the means and standard deviations of both variables), and the accuracy of these predictions depends on the strength of the association between *X* and *Y*.

Multiple Regression Introduction

I know this chapter seems like it will never end, but we're almost done. The correlation and regression topics we just learned all involved two variables: *X* and *Y*. What is the relationship between this one independent variable and this one dependent variable? Many interesting research questions can be explored using just one *X* variable and one *Y* variable. But why stop there? Why not use, oh I don't know, two independent variables? (Note: We'll always have just one dependent variable.) Maybe we would find a stronger relationship if we used two independent variables. Well, we can do that. And why stop at two? Why not use three? No problem. Or four? Can do. Or five? Slow down. Let's just use two independent variables for now.

Multiple Regression Basics

Just for fun, let's take a stroll down memory lane and examine the simple, bivariate linear regression equation.

$$Y' = a + bX$$

How do we turn this into a multiple regression equation, capable of using scores on two independent variables to predict *Y*? We'll just have to add a second *X* to the equation. And, of course, this new variable will need it's own regression weight.

$$Y' = a + b_1X_1 + b_2X_2$$

Just like simple regression, there is just a single *y*-intercept. So no change there. What is different is that each independent variable gets its own regression coefficient, which we will call a partial regres-

sion coefficient. These partial regression coefficients weigh each predictor. Bigger partial regression coefficients mean greater weights.

Let's explore a multiple regression equation with a data example. A regression of Y on X_1 and X_2 results in the following equation: $Y' = -6 + 6.1X_1 + 2.5X_2$. As you can see from this equation, the coefficients are -6 for a, 6.1 for b_1, and 2.5 for b_2. Listed below are the scores on X_1 and X_2 for this sample. (I have scores on Y too, but I'll keep those hidden.) When we apply these predictor scores to the regression equation, we compute predicted Y scores for each person.

Person	X_1	X_2	Y'
John	7	10	61.7
Molly	9	10	73.9
Neil	9	20	98.9
Chris	5	16	64.5
Jordan	6	11	58.1

Thus, computing Y' for each person in multiple regression is just a simple algebraic exercise. It's not much more complicated for equations with more than two predictors. Just a little more algebra. Five predictors? No problem. Just a regression equation with an a and five partial regression coefficients. Plug in scores on the five variables and solve. No surprises.

We need an index of the strength of the relationship in multiple regression. Some sort of a multiple correlation. That sounds like a good name. The symbol for multiple correlation is R (capital R instead of lower case r from bivariate correlation days). R is just like r except that it ranges from 0 to 1. No negative values. It's important to understand subscripts in the multiple correlation coefficient. For our example, the multiple correlation symbol is $R_{YX_1X_2}$. For multiple correlations, always list the dependent variable (i.e., Y) first, followed by the independent variables. Sometimes people put a dot between the two, but

there's no reason to do that. Finally, double subscripting can get a little tedious, so it's not unusual to write the previous multiple correlation as R_{Y12}. In the case of our current dataset, $R_{YX_1X_2}$ is .78. How did I know it was .78? I let the computer figure it out.

Linearity

No chapter on correlation and regression would be complete without a discussion of the most important assumption of these analyses, linearity. As your patience with this chapter is wearing thin, I'll keep it brief. As before, you can find a more detailed discussion of correlation and regression assumptions in my book, *Fundamentals of Correlation and Regression*.

It should come as no surprise that something called simple linear regression has, as does correlation, an assumption of linearity. Linearity means that the rate of increase (or decrease) for scores on *Y* remains the same across the range of scores on *X*. Another way of stating linearity is that the best fitting trend line is a straight line. Yet another way of stating the linearity assumption is that the most accurate summary of the observed relationship between *X* and *Y* is also the simplest: Higher scores on *X* are associated with higher scores on *Y* (or lower, if it's a negative relationship). Contrast that statement with the following: Higher scores on *X* are associated with higher scores on *Y* until a certain point at which scores on *Y* no longer increase. That statement is considerably more complicated, both mathematically and grammatically.

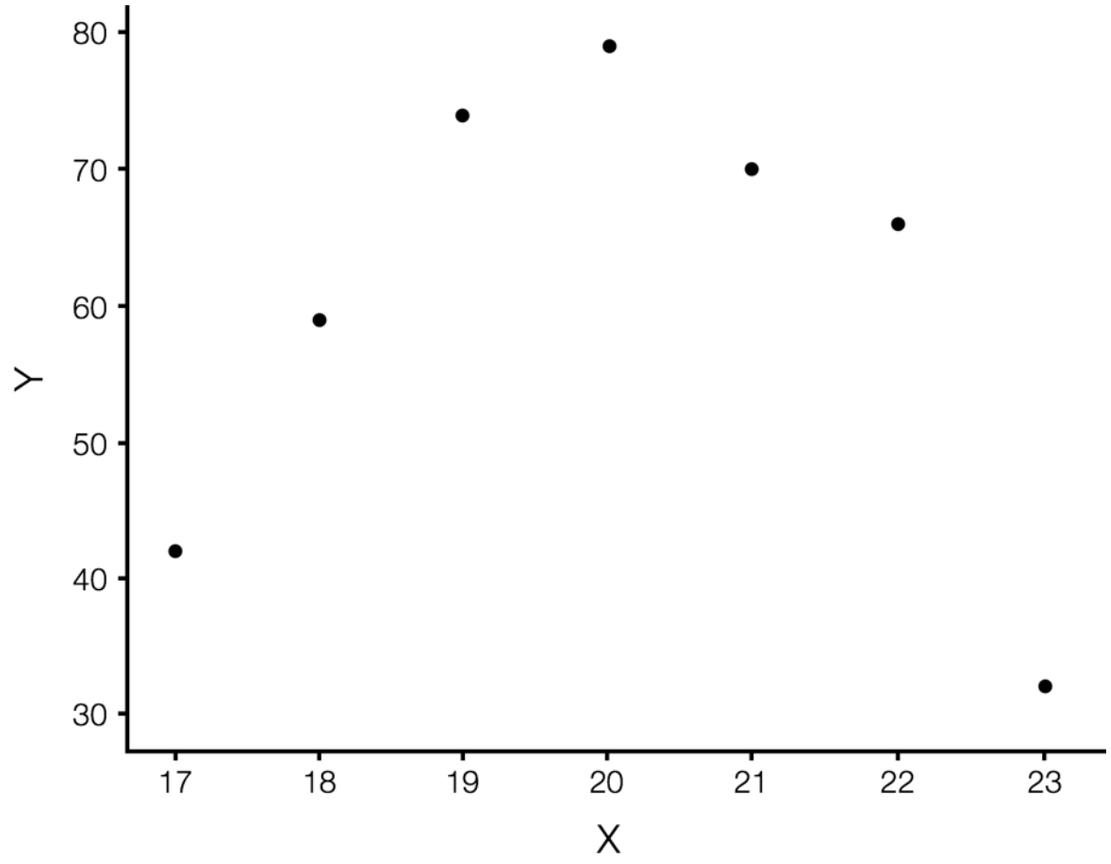
So linear regression has an assumption of linearity. What happens if this assumption is violated? If the linearity assumption is violated, a linear regression will underestimate r^2 (and *r* and *b*), and the regression equation will not accurately model the relationship between *X* and *Y*. Consider the following dataset and its scatterplot (Figure 11).

Person	X	Y
Gretchen	18	59
Steven	17	42
Jane	20	79
Mike	21	70
Brandon	23	32
Wendy	22	66
Pete	19	74

This relationship can be described as follows: Low scores on *X* are associated with low scores on *Y*, medium scores on *X* are associated high scores on *Y*, and high scores on *X* are associated with low scores on *Y* (note the complexity of this summary). Figure 11 shows a strong relationship between *X* and *Y* – that relationship just happens to be something other than a linear relationship.

As mentioned, a linear regression underestimates the strength of the relationship when the linearity assumption is violated. Back at the beginning of the chapter, we stated that for a scatterplot, the strength of the relationship is demonstrated by how close the points are to the regression line. Well, let's apply that principle to Figure 11. No matter where you draw a straight line on it, at least half of the points will have a large vertical distance between those points and the line. In

FIGURE 11 Nonlinear Trend Scatterplot



this case, a linear regression of Y on X results in an r^2 of .008 ($r = .09$). Thus, a linear regression of these data indicates an extremely weak relationship between X and Y when there is in fact a strong relationship between X and Y.

In closing, a linear model does not properly describe a non-linear relationship. That's the bad news. The good news is that there is a way to conduct a regression analysis that doesn't require a linearity assumption. The even better news (for you) is that non-linear regression analysis is far beyond the scope of this book.

Closing Thoughts on Regression

At the beginning of this chapter, we mentioned that correlation and regression have much in common. They both express the relationship between two variables through various indices of association (r for correlation, r^2 and b for regression). Given the same data, their significance tests produce the same result (the F test of R^2 yields the same result as the two-tailed t test of r). Beyond these similarities, regression analysis offers a regression equation, which we can use to predict scores on Y given scores on X. In multiple regression, the regression equation shows the unique relationship between each independent variable and Y.

There is one other difference between correlation and regression, and it's more of a theoretical difference. Correlation describes how well two variables are related and nothing else. Regression treats the dependent variable as something to explained. To be specific, regression analysis breaks the dependent variable into two parts: a part related to X (or the various independent variables) and a part unrelated to X. Going back to our Venn diagram (Figure 9), the part of Y overlapped by X (i.e., the yellow part) is the part of Y explained by X; the part of Y not covered by X (i.e., the white part) is the part not explained by X.

All very obvious, you say. Let's add a bit more to this. Remember the residual ($Y - Y'$)? It reflects the error of prediction. The residual quantifies, for each person, the lack of relationship X has with Y. How do we know this? Well, if X was perfectly related to Y, then the residual scores would be zero for each person – the predicted Y (i.e., Y') would be a perfect match to the actual Y. So there's the unrelated part. What about the related part? That's just predicted Y (i.e., Y'). How do we know this? For the same reason – if X and Y are perfectly related to each other, then all scores on Y' would be a perfect match to Y (once again, giving us residual scores of zero for all people in the dataset). Moreover, predicted Y comes from the regression equation ($Y' = bX + a$), which is the equation used to weight scores on X to obtain the best possible prediction of Y.

Let's write it out. In regression analysis, scores on Y are divided into a part related to X (Y') and a part unrelated to X ($Y - Y'$). In equation form, it looks like this:

$$Y = Y' + (Y - Y')$$

There, that's really the end of the chapter. Sorry it was so long.

Classical Test Theory

4

It may not look like much,
but it has it where it counts.

Introduction

To determine if we have a good measurement device, we need a definition of *good*. There are a number of definitions of good measurement. Classical Test Theory, Generalizability Theory, and Item Response Theory are all theories of measurement that define good measurement in slightly different ways. This chapter covers Classical Test Theory. I'd like to talk about the other two, but the title of the chapter won't allow me to do it.

Measurement Errors

Before we talk about specific theories of measurement, we need to have a serious discussion about error. There are many kinds of errors, guessing, cheating, and accidentally writing down the wrong answer when you know the right answer, to name a few. Errors include all situations in which someone earns undeserved points or fails to earn points that they do deserve. Thus, being good at taking multiple choice tests (called test wiseness) is an error of measurement because the test taker will get more points than he deserves. Being a little sick on the day of the test is an error of measurement because presumably the test taker will underperform and fail to earn points that she deserves. Poorly written test items are another source of error because they cause confusion among test takers. This confusion causes different test takers to interpret the same question in different ways, and thus fail to respond in the way the test giver intended. You can think of these errors of measurement in two ways: temporary (e.g., guessing, accidentally writing down the wrong answer when one knows the right one) or repeatable (e.g., test wiseness, poorly written items).

Measurement Theories

Theories of measurement explain: (a) how people obtain the scores that they receive on a test and (b) help us evaluate the quality of this measurement. We'll address the first issue first. You can think of it in two ways: Why do different people obtain different scores on a test? Or: Why does the same person receive different scores when he takes the same test multiple times?

Classical Test Theory

Classical Test Theory (CTT) is the oldest theory of measurement; its development dates to the work of Charles Spearman in his 1904 paper. The philosophy of CTT is that a test score can be understood as being composed of a stable component and a random component. That's it. That's the whole theory. Of course, the implications of this simple model are numerous.

So a test score is composed of a stable component and a random component. As you might guess, the stable part stays the same every time a given person takes the test, whereas the random part doesn't. The test score is called observed score and is given the symbol X. The stable component is called true score (with symbol T), and the random component is called error (with symbol e). For reasons that will be discussed later, both the term *true score* and *error* are unfortunate names in that they suggest too much; always remember them as describing a stable component and a random component.

Using these symbols, we can specify the CTT model for a test score in equation form.

$$X_{it} = T_i + e_{it}$$

That's not so bad. The subscripts in the equation are for a given person (i) at a given time (t). So, X_{it}, the observed score, is just the score that a person gets on the test. If the test is the ACT and per-

son *i* gets a 22 at time *t*, then X_{it} for that person at that time is 22. That's the observed score. Continuing with the model, we see that a given person's observed score at a given time (X_{it}) is a function of his or her true score (T_i) and his or her error score at that time (e_{it}). Right away we note that observed scores and error scores are specific to a given time, meaning that they change over time, whereas true scores are not. So for a given person, the true score is a constant; the reason that the observed score changes over time is because the error score changes over repeated measurements. Stated another way, the reason that a given person gets two different observed scores when measured at two different times is because of the error score in the CTT model. Had the error score been zero each time (or even the same score each time), this person would have received the same observed score both times. Of course, the reason why two different people get two different observed scores is because, well, they are different people (the *i* subscript) who have different true scores (the error scores are probably different as well).

Just to make sure we have it stated clearly, here again are the two fundamental tenets of CTT:

1. The true score (*T*) for a given person is constant across measurements.
2. The *e* score is random.

Now for an example.

Person	*X*	*T*	*e*
Holly	15	16	-1
John	16	14	2
Hans	6	6	0

Consider the case of Holly. Holly gets a score of 15 on a multiple choice vocabulary test; her observed score is 15. Holly knew the answer to another question, but accidentally circled "C" when she meant to circle "B". That's an error of measure-

ment. A random error, to be specific. According to CTT ($X_{it} = T_i + e_{it}$), we could describe her score as 15 = 16 + -1. That's a true score of 16 and an error score of -1. Now John takes the same test and scores a 16. But John guessed correctly on two questions he really didn't know. His observed score of 16 breaks down to a true score of 14 and an error score of +2. It should be clear by now that errors can raise or lower a person's observed score. The final case is Hans. Hans scores a 6 on the test. He never guessed correctly and he never missed any questions that he actually knew. His score of 6 breaks down into a true score of 6 and an error score of 0. So the final point is that the observed score equals the true score when there are no random errors of measurement.

Finer Points of CTT

CTT is simple so far, but we have to gum it up with a few details. As mentioned, *e* is defined to be a random component. Because we are not trying to measure randomness, any random component of a test score must be considered an error of measurement. Thus, *e* refers to random error. A better name for the *e* term would be *random error* (and we could use *re* as the symbol). But that ship has sailed. It's too late for renaming the parts of the CTT model. Back to *e*. There are plenty of sources of random errors (e.g., guessing, accidentally marking the wrong answer on the answer sheet), but there are plenty of non-random errors as well. More on those later.

On to true scores. First off, *true score* is a terrible name. The name implies that it represents "the truth" about the person. As in the person's true standing on the construct. Nothing could be further from the truth. The true score is defined as

the part of the observed score that is not e (literally, $T_{it} = X_i - e_{it}$), meaning that the true score is the part of the observed score that is not random error. To restate, the true score is just what is left after removing e, (random error) from the observed score. The definition of true score is one of those, "here's what it's not" definitions. As in, "Clouds are the part of the sky that are not blue." These definitions do not really tell us what something is, just what it is not. So we'll have to figure out what true scores are ourselves.

We can solve the mystery that is the true score by identifying every factor that influences a test score. There's construct standing. We like that. That's what we're trying to measure. Too bad that's not the only part. There's random error. Not a big fan of that. The only other thing left is some kind of error that is non-random (or systematic) error. These are errors that occur in the same way over time. We could write this out as follows:

Test Score = Construct Standing + Systematic Error + Random Error

Here's the bad news. The CTT model is $X_{it} = T_i + e_{it}$. And because CTT defines e to be a random variable, e can only include random errors. Guess where the non-random errors go? There's only one other spot, and it's T. Thus, true scores include standing on the construct *and* non-random (or systematic) errors. That is the major limitation (or weakness) of CTT.

To summarize, CTT says a person's test score (X) is the sum of a random error component (e) and a component that is not random error (T), which includes other types of errors. This appears to be a serious flaw. The good news is that even though the CTT model seems almost absurdly simplistic, it is also very useful.

A further discussion of e, which CTT assumes is a random variable, is in order. What is a random variable? A random variable does not have a sys-

tematic relationship with any other variable (or even with itself) over multiple measurements (i.e., scores on a random variable measured twice will not correlate). A good example is the roll of a die. All six values are equally likely to occur. Moreover, there is no relationship between the person rolling the die and the score they obtain. In fact, the score on the die is unrelated to anything. But over time the high scores should offset the low scores, resulting in a predictable mean score of 3.5 (1+2+3+4+5+6 divided by 6 = 3.5). The random error component of CTT behaves in the same manner. The positive errors (like guessing correctly) should cancel out the negative errors (like writing down the wrong answer to a question that you actually knew) resulting in an average error score of zero. Thus, if the same person were to take the same test repeatedly (say 100 times) with no memory of the previous times (just go with me on this one), the error scores should cancel out and the average of the observed scores equals the true score. (Remember, the true score stays the same. The observed score only changes because the error part is changing randomly.)

Problems with CTT

The problems with CTT concern its two basic principles: true scores are constant over time for a given person and *e* is a random variable. As to the first, very few constructs remain perfectly constant over time for a given person. Some constructs can change rather quickly (for example, mood), whereas others change slowly (e.g., adult intelligence). Let's take a simple one: weight. Does the weight of an adult human stay perfectly constant over the span of a year? A month? A week? A day? Of course not. A person will not likely weigh the same from one year to the next, and the difference is not due to any random factor. There is a real change in the person's standing on the construct. Are there problems with the fact that this assumption is not supported? The answer is "not much"

as long as: (a) the construct does not change rapidly and (b) the time between measurements is not long. Even small problems are still problems, though. Changes in the standing on the construct will change the observed score, which may make it appear that the test is not working well (more on this later).

The second tenet of CTT, that *e* is a random variable, is a far greater problem. As we discussed earlier, errors can be grouped into two broad categories: those likely to recur over time (non-random, like test-wiseness) and those that are one time events (random, like guessing). The fact that *e* includes only random errors means that the non-random (or systematic) errors must be assigned to the only other term in the equation: *T* (the true score). Thus, a person's true score contains systematic error. Thus, a person's true score in CTT does not simply indicate his or her standing on the construct, it also contains his or her net standing on the construct and the collection of systematic errors. In summary, don't think of true score as meaning "the truth," or a person's score free from errors of measurement. It is nothing more than a person's score free from *random errors* of measurement (as mentioned above, $T_i = X_{it} - e_{it}$).

Reliability

Now we get to a new major issue: reliability. Reliability (symbolized as r_{XX}) can be thought of as consistency. If you want it defined in three words, it's consistency of scores. To illustrate, consider the example of the scale measuring weight. The scale is a test (measurement device). Obviously, the scale is different from the tests that we traditionally use in psychology, but all of the principles are the same. Now what if you get on the scale and it says you weigh 160 pounds. Then you get off for a second and get back on again and it says 160 pounds again. So far, everything appears normal – the scale has given you the same score both times you stepped on it and clearly your

weight did not change between the two measurements. That's consistent scoring. If a test gave a bunch of inconsistent scores (imagine stepping on the scale and getting weights of 160, 76, 123, 452), then you would say, "What's wrong with this scale?" Inconsistent scores (when they should be consistent) are the result of random error. If you want to see an equation for computing reliability, first let's remember that in CTT the observed score equals the true score plus random error ($X_{it} = T_i + e_{it}$). For unimportant reasons, that equation can be written in terms of variance in which the variance of the observed scores equals the variance of the true scores plus the variance of the error scores: $S_X^2 = S_T^2 + S_e^2$. This equation can be stated as: Across a group of people, differences in observed scores (i.e., variance) is the sum of differences in true scores plus differences in e scores. Because e is a random variable, more e in people's scores means more e variance. Using this variance version, CTT models reliability as:

$$r_{XX} = \frac{S_T^2}{S_X^2}$$

Thus, CTT defines reliability as the ratio of true score variance to observed score variance. When all scores are measured without random error, the e term in $X = T + e$ becomes zero ($X = T + 0$), and thus, the variance of e will be zero ($S_X^2 = S_T^2 + 0$) – which makes the true score variance equal the observed score variance. In such a case, the reliability is 1.0. Not sure? Let's say the observed variance is 10 units. If there are no random errors of measurement, then $X = T + 0$ for everyone, and $S_X^2 = S_T^2 + 0$ ($10 = 10 + 0$). When you plug it into the equation above, $r_{XX} = 10/10$, which is 1.0. In summary, when you measure something without any random errors of measurement, $r_{XX} = 1.0$.

What about the opposite case, what if all you measure are random errors (think back to our die rolling test)? In such a case, $X = T + e$ would mean that whatever T is, it is the same for everyone.

And if everyone has the same T, then the variance of T is zero. Thus, all of the variance in X is due to variance in e. Plug this into the reliability equation and you find that $r_{XX} = 0$. So, CTT reliability tells us how much random error we are measuring. Or, to be more accurate, how much random error we are not measuring. I like to think of reliability as describing freedom from random error. This can be illustrated with a simple rewrite of our reliability equation.

$$r_{XX} = 1 - \frac{S_e^2}{S_X^2}$$

You can interpret this as follows. The reliability coefficient tells you the percent of observed score variance that is not random error. See, more random error will lead to a lower reliability coefficient and less random error will lead to a high reliability coefficient.

Now let's talk about the symbol for reliability: r_{XX}. Notice how we're using the correlation symbol r, but we are subscripting the same variable twice. It's almost as if we're saying that reliability is the correlation between variable X and variable X. In the next chapter, you'll see that we'll be doing just that.

Finally, recall that we defined the reliability coefficient as indicating a test's freedom from random error. Why couldn't we make our definition shorter and simply say that reliability is freedom from error (as in all errors, both random and systematic)? The answer goes back to the second principle of CTT. Namely, that the e term is a random variable. As we discussed, there are many non-random errors, and they become part of the true score (T). What does all of this mean to us? It means that reliability doesn't really tell us as much as we want. This is the major limitation of CTT. We may want an index of how much error (of all types) is involved in our test, but we aren't going to get it. The best that CTT has to offer is a

coefficient telling us how much random error is measured by our test.

One last detail. There is another statistic called the reliability index. The reliability index is computed as the square root of the reliability coefficient. The reliability index is almost never mentioned, but in case you see it, realize that it is not the same as the reliability coefficient (although it is very easy to transform one to the other).

A Reliable Test Is Not Necessarily a Good Test

I hope that it is clear by now that an unreliable test is bad. It measures nothing but random error. But is the converse true? Is a reliable test a good test? The answer is a solid maybe. To explain, let's substitute the word valid for good. Is a reliable test a valid test? Maybe. In Chapter 10 we'll learn that validity means that we have evidence to support the interpretations we draw from test scores. What are these interpretations? If the test measures schizophrenia and someone has a high score, the interpretation would be that the person is suffering from schizophrenia. If the test is the ACT and someone has a low score, the interpretation would be that this person will not succeed in college. Thus, there are many possible interpretations that we can draw from test scores. We'll have evidence to support only some of these possible interpretations. For example, I'm not aware of any evidence to support the following interpretation: People with high ACT scores are likely to suffer from schizophrenia. Now back to reliability. A reliable test (like the ACT) may be valid for some purposes (like predicting college performance) but not for other purposes (like identifying various personality disorders). It all depends on what we have evidence to support. Of course, there are many reliable tests for which we don't have any validity evidence for any purpose. So to sum up, an unreliable test cannot be valid for any purpose (it measures random error which

is uninterpretable as being anything other than randomness). A reliable test may be valid for a given purpose, depending on where the evidence lies.

Estimating Reliability

We can never know the actual reliability of a test, but we can make a decent estimate.

Some of the time.

A Brief Review of Classical Test Theory

In the previous chapter we learned that CTT defines a test score (X) as the sum of a true score (T) and error (e) with the simple equation $X_{it} = T_i + e_{it}$. Also, the variance of each term works in the same fashion: $S_X^2 = S_T^2 + S_e^2$. Reliability is defined as the ratio of true score variance to observed score variance ($r_{XX} = S_T^2 / S_X^2$). Thus, to compute the reliability coefficient (r_{XX}), all we need to do is gather data from a group of people, compute their observed scores, compute their true scores, compute the variance of both terms, and divide. Piece of cake. Easy as pie. Anyone getting hungry?

A cloud forms on the horizon. Computing the observed scores is easy enough, but how do we compute the true scores for each person? CTT defines the true score as being the part of the observed score that is not random error. OK, so we'll need to know the observed score (easy enough) and the error score for each person. Big problem: We can never know the error score for a person. Which means we can't compute the true scores and can't compute the reliability coefficient. The whole house of cards comes crashing down. (Side note: All of the examples involving true scores and error scores in Chapter 4 were constructed to illustrate CTT principles. None were real data because none could be real data.) Let me assure you that this whole thing wasn't just a waste of time. We will continue to use these CTT terms and principles.

There is a second definition of CTT reliability which states that reliability can be computed as the correlation between parallel tests. And parallel tests are two forms of a test that:

1. Measures the same construct(s) with different items.
2. Are equal in quality.

Equal in quality is further defined as having the same means, standard deviations, and correlations

with external variables. Which pretty much means equal in every way a person can imagine. And I can imagine a lot. Now we're not saying that both versions of the test have to be perfect, just equal in quality. If one version is bad, the other version must be equally bad for it to be a parallel test.

Now how does this new definition (correlation between parallel tests) fit with the old one (ratio of true score variance to observed score variance)? Are we just making up something to get ourselves out of a tight spot? Not at all. Both definitions are interchangeable. Here's how. First we need to remember that a random variable will not correlate with anything. That's the nature of randomness. Randomness is inherently unsystematic – no patterns, no trends, nothing. And something unsystematic can't have a systematic relationship with anything. That means a random variable will have a zero correlation with any other variable. Think about our die rolling test. If a large group of people (say, 1000) all rolled a die and then took an intelligence test, would we expect to find a non-zero correlation between our two tests? Do smarter people roll higher scores? Or could it be that it is the slower people who roll the higher scores? Of course not. The correlation would be zero. That's the way it is with random data.

Back to the definitions. You know from Chapter 4 that when we measure nothing but random error, there is no true score variance, and thus, the ratio of true score variance to observed score variance is zero. When our measurement has no random error, all variance is true score variance and the ratio is 1.0. Given what we know about correlations and random error, we can safely say that when our parallel tests are full of random error, the resultant correlation will be zero. When our parallel tests have no random error, the correlation between them will be perfect. In short, we get the same result from both definitions of reliability. I hope this has put your mind at ease over the matter.

Reliability is a correlation between parallel tests. To compute reliability we'll give a group of people one version of the test (which we'll call Form A). When they finish it, we'll give them the other version of the test (Form B) and correlate the scores. All very simple. If the person with the highest score on Form A also has the highest score on Form B, and the person with the second highest score on Form A also has the second highest score on Form B (and so on down the line), then we'll likely get the perfect correlation and we'll conclude that our test has perfect reliability ($r_{XX} = 1.0$). (Side note: Now does the symbol for reliability, r_{XX}, make sense? We literally are correlating a test with itself.)

It should be obvious that it will be difficult to actually have parallel tests. In fact, we'll dispense with any optimism and say that parallel tests are a hypothetical entity that will never be found in practice. And although we may seem stuck, we're now close enough to reality that we can make a small jump from the real world of data to the fantasy world of parallel tests. Here's how: We'll just cheat on the definition of parallel tests. Unfortunately, cheating comes with a price. We'll never really know the real reliability of our test, we'll just get an *estimate* of its reliability. This estimate might be high quality, or it may be low quality. A high quality estimate, although not spot on, can still be very useful.

Alternate Forms Reliability

When I mentioned that parallel forms do not actually exist, you might have said something like, "Let's just make two versions of a test and pretend that they are parallel!" First off, I salute your enthusiasm. Second, we can do that, but let's be honest about it and call these tests *alternate* forms instead of parallel forms. Because you know that we'll never be able to come up with two different versions of a test that satisfy part two of the parallel test definition (equal in quality). With alter-

nate forms reliability we'll do our best to make them as close as possible in quality, but no matter how hard we try, one of them will be a little more difficult than the other and/or correlate with some other variable (like college GPA) a little more strongly than the other.

Other than that issue, alternate forms reliability works just like the parallel tests way of computing reliability. First, we'll give Form A to a group of people. Next, the same people will complete Form B. Finally, we'll correlate the scores. It's just that simple. We'll give this correlation a special name (in addition to the more generic *reliability coefficient*). We'll call it the coefficient of equivalence. Sounds important. The only reason to ever call it by its special name is that when we tell people that our coefficient of equivalence is .8 (or what have you), they know that we estimated reliability with an alternate forms study.

From a theoretical level, the only problem with alternate forms reliability is the obvious issue: What if our tests are not anywhere close to equal in quality? That is, what if Form A contains a bunch of well written questions but Form B is full of a bunch of terrible and confusing questions? In such a situation, it should be obvious that our correlation will be affected. (Remember, randomness doesn't correlate, and Form B is full of random error.) Whether it's too high or too low is a matter of perspective. It will underestimate the reliability of Form A (the good test) and will overestimate the reliability of Form B (the bad one). Because we don't know in reality which test is the good one, there is no way for us to know whether the reliability is over or underestimated. (Just to take an extreme example of this problem in action, let's say that Form B is complete random error. Reliability is zero for From B. In this case, Form A could have perfect reliability, and the correlation between the two would still come out

to be zero. Sure, a reliability estimate of zero is correct for Form B, but not for Form A. For Form A, it's a severe underestimate.)

This brings up other unpleasant issues in alternate forms reliability. Namely, we have to make two complete versions of our test. It's hard enough to develop one good test, let alone two. And as discussed above, developing one and a half good ones is not enough. Finally, imagine you are one of the people in our alternate forms reliability study. You finish taking a test – maybe it was long and exhausting. Now, we ask you to take another test, just like the first one. Not a pleasant prospect. Even if you decide to participate, fatigue may affect your performance and confound our results. What if we gave our test takers a break between testing sessions? That's a good idea, one that will be addressed later.

Split-Half (or Internal Consistency) Reliability

I think we're in agreement that developing two different versions of a test is twice as much work as we want to do. There must be another way to make this work out without doing any real work. What if we took just one version of a test and split it into two equal halves? That might work. We could give people a score on the first half and another score on the second half, then correlate the scores. We'll let each half of the test serve as the two forms for our parallel tests. That might work.

Here is how we do the split-half reliability study (Interactive 1 demonstrates this process). We give our test to a group of people (remember, we don't have an alternate form – we only have one version of our test), when they finish the test, we say "goodbye," we then split the test questions into two groups, compute a total score for each half of the test, and correlate the scores. Just like

always, we want a strong correlation (which we'll get if the person with the highest score on the first half of the test also has the highest score on the second half, and so on…). We'll call this correlation the coefficient of internal consistency. (Side note: Split-half reliability studies are often called internal consistency reliability studies for obvious reasons – in these studies we are examining the consistency of scores from one part of the test with the other.)

INTERACTIVE 1 Split-Half Reliability Illustrated

Now let's do an example. Let's say we have a test with four questions and we give our test to a sample of six people. This test is a multiple choice test with four options. We've scored the data in the normal fashion so that a correct answer is a 1 and an incorrect answer is a 0. Here are the data:

Person	Item 1	Item 2	Item 3	Item 4
Augustus	1	0	0	0
Ellis	1	1	1	0
Cedric	1	1	1	1
Dennis	1	1	0	0
Jimmy	0	1	0	1
Randy	0	0	0	0

Now if this was regular test situation, we would assign everyone a total score (Augustus would have a 1, Ellis would have a 3, Cedric would have a 4, etc.). But this is not a regular situation, we need to divide the test in half and compute two scores for each person – one for each

half. We'll do a simple first half (Items 1 and 2) versus second half (Items 3 and 4) split.

Person	Item 1	Item 2	Item 3	Item 4	First Half	Second Half
Augustus	1	0	0	0	1	0
Ellis	1	1	1	0	2	1
Cedric	1	1	1	1	2	2
Dennis	1	1	0	0	2	0
Jimmy	0	1	0	1	1	1
Randy	0	0	0	0	0	0

At this point, all we have to do is correlate the scores. In our case the correlation is .50. That's it. That's the coefficient of internal consistency. That's the whole split-half reliability study. Pretty simple. Just give the test like you would any other test, score the test in two parts, and correlate the scores. Lather, rinse, repeat. What could be simpler?

Well, here's the problem. Three of them, actually. Ready? Here we go. The first problem is that when we compute a split-half reliability, we are getting a reliability for a half-length test. A half-length test is one that is only half as long as the actual test. Think about it this way, how many items are on our test? Four. And if we did an alternate forms reliability, how many items would be on Form A? Four. And Form B? Four. And in an alternate forms reliability study, we correlate our four item total score on Form A with our four item total score on Form B to get the reliability coefficient. But in our split-half reliability study, how many items are on the first half? Two. And the second half? Two. So the reliability coefficient in our split-half reliability study is based on the correlation between two *two-item* subtests. In short, in a split-half reliability study, we are getting a reliability estimate for a test that is only half as long as the actual test. Here's why this half-length test thing is a problem: Other factors being equal,

longer tests are more reliable (and shorter tests are less reliable). Thus, our split-half estimate of reliability will be lower than the actual reliability. Remember, this wasn't a problem with alternate forms reliability because we never cut that test in half – we created two complete versions of the test.

Now that we know what the first problem is (and before I tell you the solution to it), allow me to digress and explain why longer tests are more reliable. First, recall that reliability means freedom from random error. Playing our "what if" game, what if we had a one-item test? Say it's a multiple choice math test. One question, that's all. Your score is either a 0 or 100. No chance for partial credit. What if you know the answer but accidentally write down the wrong answer? That's an error of the random kind. What if you don't know the answer but you guess correctly? Another random error. In both cases, the impact of these random errors is huge: Your score goes from the highest possible to the lowest possible value (or vice versa). Now what if we had a two-question test and you make a random error on one of the questions? In this case, the magnitude of the error (50 points) is not as great. Moreover, there is a chance that we'll make two complimentary random errors that reduce their ultimate impact to zero. Huh? Let's say that we don't know Item 1 but guess correctly (we get 50 points that we don't deserve) and we do know Item 2 but write down the wrong answer (we lose 50 points that we should have earned). Our final score is 50, which is exactly what it would have been if we hadn't made any random errors. Essentially, the positive error was offset by the negative error. Now I'll be the first to admit that this offsetting error situation in a two-item test is ridiculously unlikely, but it can happen for a two-item test and could never happen in a one-item test. The longer the test, the less the impact of a single random error (on a 100-item test, a single random error is worth only a single point)

and the increased likelihood that the random errors will cancel out for a given person. To restate, other things being equal, longer tests are more reliable. What are these "other things?" The "other things" refer to the quality of the questions. If we have a ten question test composed of very good items and add to it another ten very bad questions, our test, though longer, is not more reliable.

Now that we have that straight, here's the solution to the half-length test problem. A simple formula called the Spearman-Brown Prophesy formula (coolest equation name ever) will tell us what the reliability for a test would be given our current reliability and any changes in test length. In other words, it's a "what if" formula. And it is quite obvious that it is the perfect equation for our half-length test problem. The equation is given below.

$$r_{XX}SB = \frac{k \cdot r_{XX}}{1 + (k-1)r_{XX}}$$

Where:
r_{XX} is the reliability estimate computed in our study.
$r_{XX}SB$ is the new corrected reliability estimate
k is the factor by which we are lengthening or shortening the test.

It's important to note that k is not the number of items on the test. k is a ratio. If we are doubling the length of the test, then k is 2. If we are tripling the length of our test, k is 3. If we are cutting our test in half, k is .5. We can compute any of these *what if* scenarios with Spearman-Brown, but the one that is most relevant to us is the doubling scenario. Remember that a split-half reliability study gives us a reliability estimate for a half-length test, which is lower than the real reliability of the test. So we want to correct for the half-length problem. We want to raise it up to its original full length. To get one half back to one, we double it. Thus, k will be 2. When we plug our .50 reliability coefficient (obtained from previous dataset) into Spearman-

Brown with a *k* of 2, we get a corrected value of .67. This .67 estimate should be closer to the real reliability.

Just to remind you that Spearman-Brown can be used for other purposes, let's say that we got a coefficient of equivalence (remember that?) of .7 and we want to know what our reliability would be if we tripled the length of each version of the test. A quick calculation tells us that the new reliability would be .88. A mighty fine number. Prompted by this pleasing forecast, we get to work writing a bunch of new items.

Finally, the "other things" qualification applies to Spearman-Brown projections as well. In the previous example where we explored what would happen if our test length was tripled, we would only obtain this new reliability of .88 if the new items that we write were equal in quality to the items that we had already written. In short, we can't add a bunch of garbage to our test and expect good results simply because the test is longer.

Problem number one took forever. I'm drained. Here's problem number two. For a split-half reliability study, we split the test into two halves. In our example from above, we split the test in a first-half versus second-half fashion. That's not the only way we could have split it. We could have split it in an odd (Items 1 and 3) versus even (Items 2 and 4) fashion. In fact, let's do that.

Person	Item 1	Item 2	Item 3	Item 4	Odd	Even
Augustus	1	0	0	0	1	0
Ellis	1	1	1	0	2	1
Cedric	1	1	1	1	2	2
Dennis	1	1	0	0	1	1
Jimmy	0	1	0	1	0	2
Randy	0	0	0	0	0	0

When we correlate the scores from an odd/even split we get a correlation of .25. Wait a second, the first time we did this, we obtained an uncorrected correlation of .50. Now, it's .25. The original data are the same, the only thing we changed is how we split the test. That's a big change over what seems to be a minor detail. You know what, we could split this test another way: Items 1 and 4 versus Items 2 and 3. Let's see what happens when we do that.

Person	Item 1	Item 2	Item 3	Item 4	1 & 4	2 & 3
Augustus	1	0	0	0	1	0
Ellis	1	1	1	0	1	2
Cedric	1	1	1	1	2	2
Dennis	1	1	0	0	1	1
Jimmy	0	1	0	1	1	1
Randy	0	0	0	0	0	0

When we correlate the splits this time, we get a .70 correlation. That's three different splits, three different correlations. Which one is the correct one? Do we chose the highest? The lowest? The middle-est? Some kind of average? Do we just go with whichever one we got the first time? I hope that you can see that if we do just one split, we might get an exceptionally good split (r_{XX} = .70), a bad split (r_{XX} = .25), or something in between (r_{XX} = .50). In our example with a four question test, there were only three ways to split it, and thus, only three possible reliability coefficients. What if our test had 20 items? In that case, there would be over 92,378 possible splits, yielding 92,378 different reliability coefficients. Would our chances of getting one of those extreme splits increase? I honestly don't know, but let's not chance it.

So what's the solution to the "how do we make the split problem?" The answer is to do them all. Every split. Do every split, compute each

of the reliability coefficients, and then compute the average of them. As a final bonus, we'll use the Spearman-Brown correction on this average correlation. Before you say, "That sounds like work," there is a formula which will do all of this work for us. It's called coefficient alpha (or Cronbach's alpha), and it is remarkably simple.

$$\alpha = \frac{k}{(k-1)}\left[1 - \frac{\sum S^2_{items}}{S^2_{total}}\right]$$

Where:
k is the number of items on the test (a different k from before)
S^2_{items} is the variance of each test item
S^2_{total} is the variance of the total score

I am amazed at just how much work coefficient alpha does. Everything mentioned in the previous paragraph using just three terms, k and two variance terms. As mentioned, k in coefficient alpha refers to the number of items on the test, which was not the case in Spearman-Brown. Same letter. Different meanings. Don't get mad at me – I don't make this stuff up.

So, to compute coefficient alpha, we compute: the variance of each item (which we then sum across items), the total score on the test for each person, the variance of the total scores, and the number of items. Plug in and compute. The answer that we get will be (almost) the same as the number we would have obtained had we computed all possible splits, computed the correlations from all possible splits, averaged this correlation, and corrected this average with the Spearman-Brown prophesy formula. All that work in one simple equation.

Thus, with coefficient alpha we no longer have to actually split our test. We just compute the variance of each item and the variance of the total score. Much easier. And it solves our "How do we make the split?" problem. At this point, you may

More Thoughts on Coefficient Alpha

When I introduced coefficient alpha I stated that coefficient alpha equals the mean of all possible split-half correlations with a Spearman-Brown correction. Well, that's not entirely true. They only work out to be the same if all of the items have the same variance, which they won't. But even though it is not exactly the same, it's very close.

Another way to think about coefficient alpha is that alpha is the mean correlation between each pair of items (with a Spearman-Brown correction thrown in at the end). This means you correlate every item with every other item on the test – that's a pile of correlations. Then take the average of those correlations. And then give the average correlation the Spearman-Brown treatment. But

be thinking, "Why did we spend all of that time learning about split-half stuff when all we needed was coefficient alpha?" Good question. The answers are twofold. First, coefficient alpha makes more sense if you understand an old fashioned split-half reliability study. Second, there are situations in which there is so much missing data (i.e., we don't have a score for Item 12 for a given test taker), that we cannot compute coefficient alpha and must compute the old split-half reliability.

There are some other equations which do the same thing as coefficient alpha. The most popular are Kuder-Richardson 20 and Kuder-Richardson 21 (also called KR-20 and KR-21). Both of these can be best described as simpler (and in the case of KR-21, less accurate) versions of coefficient alpha. They are simpler in that they require less work to compute. In the computer age, however, the amount of work needed to compute an answer is a trivial issue, rendering both KR-20 and KR-21 quite inconsequential.

Regardless of whether we estimate internal consistency with an old fashioned split-half study,

coefficient alpha, KR-20, or KR-21, we still call the end product a coefficient of internal consistency.

Finally, we can get around to the third problem with all internal consistency estimates of reliability. All internal consistency reliability estimates will be downwardly biased if the test is multidimensional (i.e., measures more than one construct). That is, our internal consistency estimate of reliability of will be lower than it should be if our test is multidimensional. In a nice twist, we'll use this problem to our advantage when we get to item analysis (short version: we'll throw out items which cause coefficient alpha to be low). That said, if our goal is to get an unbiased estimate of reliability and we suspect that our test is multidimensional, we should not use an internal consistency strategy. How can we estimate reliability in such a situation? Use the previously mentioned method, alternate forms reliability, or use the next method, test-retest reliability.

Final Thoughts on Internal Consistency Reliability

Whether we compute internal consistency with a split-half correlation or with coefficient alpha, the general point is that a correlation among items (or halves of test) is an index of reliability.

Why? Why is it that the correlation among items on a test works as a reliability estimate? Let's answer that question with another question: Why do items fail to correlate with each other? Items fail to correlate for two reasons: These items measure too much random error, or these items measure different constructs. Remember that randomness doesn't correlate. So if any one item is full of ran-

Test-Retest Reliability

Recall that CTT reliability is defined (among other things) as a correlation between parallel tests. And parallel tests are two forms of a test that (a) measure the same construct with different items and (b) measure the construct equally well. What if we simply gave the test twice? That is, a

group of people takes the test one time and then takes the exact same test a second time? We would be using the test as its own parallel form. That could work. You might say that this is a shortcut on the alternate forms method. Instead of taking the time to develop a second form, we'll just get lazy and use the same form at both testing sessions. Of course there is a price to be paid. Will both versions of the test be equal in quality? They should since they are the exact same questions. Are we measuring the same construct with different items? No, it will be the same items. Thus, as with alternate forms reliability and split-half reliability, we are satisfying one part of the parallel tests definition, but not the other. As usual, we will end up with a biased reliability estimate.

Our basic procedure for a test-retest reliability study is as follows: (a) administer test to a group of people, (b) wait some amount of time (called the intertest interval or ITI), (c) administer the same test to the same group of people, and (d) correlate the scores from the first time they took the test with the scores from the second time they took the test. This correlation is our reliability estimate and we'll call it the coefficient of stability as it describes how stable the scores are from Time 1 to Time 2. Ideally, everyone would have the same scores both times, and, thus, the person with the highest score at Time 1 would also have the highest score at Time 2 (and the second highest person at Time 1 would be second at Time 2, and so on). The main issue with a test-retest reliability study is how long we wait between administering the test (the ITI). We could wait as little as a second (i.e., give the test a second time as soon as the finish it the first time) or as long as a lifetime (not recommended).

You may have already guessed where the bias comes from. It's due to the fact that we are administering the same set of items twice and people have a habit of remembering things if it hasn't been too long since they first saw the items. It

should be obvious that the big issue is how much time is enough for people to forget the items? The answer is no one has the faintest idea. Speaking on an anecdotal level, I recall remembering a question on a third grade standardized test that I had first seen on the second grade standardized test. That's an ITI of one year, but I hadn't forgotten it. (Of course, it was only one item of many on the test, and I didn't remember the other dozens of items. But still...)

Now let's talk about the impact of the ITI. If the ITI is short and people remember a bunch of the questions from Time 1, their scores will *likely* be more similar than they should be (bear in mind that with a test-retest reliability study, we are interested in how similar the scores are from Time 1 to Time 2). Notice that I italicized the word *likely*. It won't be that way for everyone on every type of test, but there are two important principles that make scores more similar than they should be: laziness and a desire to appear consistent.

Laziness first: If I worked for a while on a tough math problem and came up with an answer of 4 the first time I took the test, and I see the same math problem again (and remember that I answered 4 the first time), what are the chances that I'll do all of the work again? Not much. I'll just put 4 and move on without a second thought. By doing this, if I made a random error the first time, I am dooming myself to make the same error the second time. Thus, my random error has just become not so random. (Conversely, if my answer of 4 the first time was correct and I put the same answer the second time from memory, I am depriving myself of the opportunity of making a random error the second time.) My score is more stable than it should be because I recalled the question and my answer at Time 2. The correlation between scores at Time 1 and Time 2 is higher than it should be, and thus, reliability is overestimated.

The desire to appear consistent (or avoid looking wishy-washy) becomes relevant on measures

of personality or attitudes. Let's say you ask a person at Time 1 how they feel about cats, and this person likes them a little bit. Some time between Time 1 and Time 2, they have a slight change of mind and dislike cats a little. When you ask a group of such people the same question at Time 2, many people (but not all) who remember their answer from before will put the same answer again even though they feel differently about cats. Why? People are aware that society frowns on capricious, wishy-washy behavior and try to avoid showing this behavior to others. Even on a confidential personality test. The impact of this is just like before: Responses are more consistent than they should be (i.e., if people remembered nothing) and thus, reliability is overestimated.

At this point, you're thinking, "Here's an easy solution. Just use a really long ITI so that no one remembers squat." That approach will eliminate the problem of the previous paragraph, but it will create a new one. People change over time. To use CTT terminology, we would say that their true scores change over time (recall that true scores include a person's standing on the construct as well as non-random errors). But CTT assumes that true scores stay constant. Well, on a long enough timeline, everyone's true scores will change for just about any construct. The changes might be big, they might be small, but there will be changes. Because CTT assumes that true scores remain constant, any change is considered error. The correlations between scores at Time 1 and Time 2 are lowered, and the test looks less reliable than it really is. An example should help clear this up. Let's say the construct we want to measure is a person's weight in pounds. And let's say that we happen to have a perfect scale. The scale is our test and let's not ask how we know it's perfect. For the sake of this example, that's not important. So imagine that we have a group of 100 people and we weigh them at Time 1, wait 10 seconds, and weigh them again (Time 2). The correlation of

scores from Time 1 and Time 2 is our test-retest reliability coefficient. Obviously, in 10 seconds, there has been no change in anyone's real weight. Thus, their true scores have not changed. Now let's do it again, only this time the ITI is five years. Now you know that over five years, there will be some real changes in people's weight. Their standing on the construct (which is part of the true score) has changed. When we weigh them at Time 2, the observed scores have changed a lot, not because our measurement device is bad (remember, it's perfect), but because these people have changed. What happens to our correlation? It's lower. Thus, this test appears to be less reliable than it really is. Bear in mind that in reality we don't know whether our test is good. So when we do a test-retest reliability study with a long ITI and obtain a less than desirable reliability coefficient, we have to wonder: Is the test actually unreliable (rotten with random error) or is the ITI causing the low reliability coefficient through changes in the standing on the construct?

So, a short ITI is bad because it overestimates reliability and a long one is bad because it underestimates reliability. What is the right ITI? What is the ITI that is not too short or too long? It would have to be long enough that no one remembers the test, but short enough that no one has changed their standing on the construct. Well, it's complicated, and there hasn't been enough research on the issue. I only found one study, and it wasn't conclusive. Moreover, the answer to this question would depend on the type of construct measured. A mood related construct would change rapidly whereas an intelligence construct would change very slowly for an adult population. So the answer is, there is no answer. It is easy to find a number of *recommendations*. Suggestions range from two weeks (Pedhazur & Schmelkin, 1991) to six months (Nunnally, 1967) between test administrations. It appears that the sugges-

tions are based on the authors' preferences to avoid one of the two problems with intertest interval length. Some researchers believe that the drawbacks from long intervals between tests have more impact; thus they suggest a shorter interval (Pedhazur & Schmelkin) or recommend that the interval not exceed six months (Anastasi, 1982). Conversely, there are researchers who believe that the impact of a short interval is greater and who thus suggest a longer interval (Nunnally). Unfortunately, these are mere recommendations and are not based on actual data. One thing that is easy to determine is what ITIs researchers are actually using. In an analysis of 276 test-retest reliability studies, the median (and modal) intertest interval was a mere 14 days (Brown & Cromwell, 2005). (On the positive side, the mean was 66 days – much longer due to a few outliers that skewed the distribution. Another lesson in means versus medians.) Fourteen days is pretty short, essentially guaranteeing that people will recall a non-trivial amount of information, which will influence their answers in ways we discussed earlier. That said, if you are conducting your own test-retest reliability study and want to do what everyone else is doing, use an ITI of 14 days.

Knowing what we know about the relationship between ITI and the resultant estimate of test-retest reliability, we can draw some clear inferences in some cases. Let's say you see that Dr. Durden reports a coefficient of stability of .54 for a certain personality test in a study in which he waited two days between test administrations. In such a situation (short ITI), we should obtain a high correlation, but clearly we got a low one. It is pretty easy to conclude that this test has reliability problems. If a person's score isn't very consistent across two days (when he likely remembered a great deal of information from the first administration), this test is probably measuring a lot of random error. Hopefully that made sense. If not, go back a page and start over. New example of the

same issue. You run across a study from a Dr. Jack (not sure whether Jack is his first or last name – or even if it is his real name) in which the ITI is five years and the coefficient of stability is .86. Wow. Impressive. In this situation, the ITI is working against us (people have real changes over a time span of five years), and yet Dr. Jack obtained a high correlation. Must be a reliable test.

There are still more problems with a test-retest reliability, but these are easier to understand. The first problem is one of motivation. Will the test-takers be as motivated to take the test a second time? Not that this isn't a problem with parallel forms, but it is likely to be a bigger problem here. The second is that it is often difficult to get every person to take the test a second time. You may start out with 100 people at Time 1, but at Time 2 only 58 show up. Guess what your sample size is. That's right, 58. Moreover, we don't know how the results would have turned out had all 100 returned. Would the reliability have been higher? Lower? Same-er? No one knows.

Alternate Forms and Test-Retest Reliability

What if we combined the alternate forms reliability strategy with the test-retest strategy? That could be cool. We would give Form A of the test to a group of people at Time 1, wait, and then give Form B to the same group of people at Time 2. Now we've got all of the hassle of developing two forms of the test along with the irritating ITI involved with a test-retest study. We'll call the correlation a coefficient of equivalence and stability. What good would this do? Why is this needed? Clearly no one will remember questions from Time 1, because they won't see the same items at Time 2. But there will be familiarity with the item types. And familiarity breeds contempt. Rather, the practice associated with seeing the same types of items a second time could help one's performance. If this is a real concern to us, then we should

wait long enough for this familiarity to wear off. Hence the ITI. Essentially, we are tweaking the basic alternate forms strategy to eliminate a potential confound. But now we introduce a new problem. A problem with which we are familiar. What if we wait so long between Time 1 and Time 2 that the test takers' standings on the construct change? If you can remember back from two or three paragraphs ago, you know how this goes. We don't want to wait too long. How long is long enough? See previous page for the non-answer to this question.

Which Way Is the Best Way?

At this point, you may be wondering which method of estimating reliability is the best: alternate forms, split-half, test-retest, or alternate forms and test-retest. If you've been paying attention, then you know that the answer is: There is no answer. Every one of the methods are flawed in some way, rendering a biased estimate of reliability. Which is why the title of this chapter is "Estimating Reliability" and not "Computing Reliability." All of these studies will give us a number that is hopefully fairly close to the true reliability of the test, but due to flaws in the design of the studies (none are actual correlations between parallel tests), we'll never know the true answer. All we'll get is an estimate. There are some conclusions that we can draw, however. First, if you can't create two different versions of a test that are equal in quality to your eyes (of course, they will never be truly parallel), then anything involving alternate forms isn't for you. Second, if you think that there is any chance that your test is even a little multidimensional, then split-half isn't for you. Finally, if you don't have a long enough ITI (at least two months in this author's opinion – yet another unfounded recommendation), don't use a test-retest strategy.

The Strange World of Interrater Reliability

There is a final reliability term that you may hear at some point: interrater reliability. Interrater reliability is a whole different animal from the ones above. In fact it is so different, that it is not a real method for estimating reliability (according to CTT) and doesn't deserve to have the word *reliability* in its name. In fact, some researchers have advocated changing its name to *interrater consistency* (Kozlowski & Hattrup, 1992). Be we're getting far ahead of ourselves. First, what is this thing that some call interrater reliability?

Consider a typical reliability scenario: a group of people complete the same test (say, an intelligence test) twice (test-retest reliability). The test taker is the source of most of the data (scoring issues aside). That is, who provides Marla's data at Time 1? Marla. And at Time 2? Marla. And the same for Bob, and so on. Interrater reliability is different. With interrater reliability, two raters are providing ratings for each of the people. Let's say the raters are managers who rate the job performance of a group of workers (the ratees). Of course, the ratings should be based on what the worker actually do on the job but the potential for the rater to influence the final rating is much greater here than for a multiple choice test graded with a key.

Here's an example. Let's say a small group of five workers have their job performance evaluated on a 5-point scale by two managers (Rater 1 and Rater 2, no relation).

Worker	Rater 1	Rater 2
Michael	5	5
Andrew	3	2
Brian	3	3
Ian	4	5
Bill	4	4

The ratings are then correlated. We get a strong correlation when the person that Rater 1 rates highest is also rated highest by Rater 2 (and so on). This correlation is then treated as our reliability estimate. Here's what makes the interrater reliability model different from regular reliability: Who provides the score for Ian? Not Ian directly, but the managers. It is their evaluation of Ian that becomes his scores. One could argue that this really isn't all that different from every other way of estimating reliability (tests have to be scored by someone), and there is an element of truth to that. The difference is one of degree. The scorer of an intelligence test has an influence on the observed score of the test taker. The rater has much greater influence on the observed score of the ratee. Moreover, the rater may not even be aware of how she or he is influencing the rating. The net result of all of this is that with rating data, the rater is a new and greater source of random and systematic error, and any correlation between rating data will be affected by these errors.

Interrater Reliability Is/Is Not a Real Reliability. CTT says reliability is a correlation between parallel tests. Although raters assume the role of tests, they will never meet the definition of parallel tests. In short, raters are not parallel tests, which means that interrater reliability is not a real reliability according to CTT. Which is why some have stated that the correlation between ratings from two (or more) raters should be renamed interrater consistency. The correlation does tell us how consistent the ratings are, but any inferences regarding CTT reliability are ill-advised. All that said, if you want to say that these complainers are just being pedantic because although the raters are not tests (and thus, not parallel tests), they are scorers of tests, just like real tests, which makes a correlation between their scores is a real CTT reliability, there's probably no way I can talk you out

of that position. So there, I just argued it both ways. I don't feel good about it.

Interrater Agreement

The twin cousin of interrater reliability is interrater agreement. With interrater agreement, there is no pretense of a connection with CTT. So the conceptual problems listed above are problems no more. To reiterate, interrater agreement has nothing to do with CTT. I only mention it here because it contrasts with interrater reliability, which we just discussed. With interrater agreement, we are simply describing the consensus among the raters. That is, do they assign the exact same rating? To contrast, with interrater consistency we examined the pattern of the ratings (with a correlation). The ratings might be very different in an absolute sense, but as long as the person rated highest is the same for both raters, you get a strong correlation. Well that's not the case with interrater agreement. Three sample datasets illustrate the differences.

Ratee	Rater 1	Rater 2
Michael	5	5
Andrew	3	3
Brian	2	2
Ian	4	4
Bill	1	1

This first dataset should be easy to understand. Both raters assign the exact same ratings to every person. Thus, agreement is perfect – they never disagreed. If you run the correlation, you'll find that it is 1.0, perfect interrater consistency.

Ratee	Rater 1	Rater 2
Michael	5	3
Andrew	4	2
Brian	3	1
Ian	4	2
Bill	3	1

With this second dataset, things start breaking down. The raters disagree by two points on every single person rated. On a 5-point scale, two points is huge – it's the difference between saying someone is average (3) versus great (5). Interrater consistency for this dataset is perfect ($r = 1.0$). Why? The person rated highest by Rater 1 is also rated highest by Rater 2 (and so on). See Chapter 3 for more explanation on how correlations work if this feels rusty. It is this scenario which causes so many problems for measurement people. If our raters are consistently disagreeing by two points, we have a real flaw in the ratings and we shouldn't ignore it.

Ratee	Rater 1	Rater 2
Michael	5	5
Andrew	3	4
Brian	3	4
Ian	4	3
Bill	4	3

This third dataset shows us that we can have good agreement (the two raters are never more than one point apart in their ratings), but have poor interrater consistency due to a limited range in the ratings ($r = .29$). This is a rare scenario, but it can occur. It is not, however, a troubling scenario. As the second dataset demonstrated, agreement is more comprehensive (and more difficult to achieve) than consistency. As such, if we observe good interrater agreement, we likely care not that interrater consistency is poor.

How, you might ask, do we compute interrater agreement? Unfortunately, you have just asked a

complicated question. Well it's a simple question with a complicated answer. The answer depends on a number of factors including the number of raters, the type of rating they make (nominal versus interval), the number of targets rated, and whether you care about agreement at the individual target level or across all targets. The good news is that we won't worry about these issues at this juncture. Just learn the stuff above and we'll move on.

The Standard Error of Measurement

We have spent quite a bit of time discussing reliability. Maybe too much time. But what does reliability do for us? As I've said before a reliability coefficient tells us how free our measurement is from random error. We can take this knowledge and use it to estimate the role random error plays in an individual score. At this point, it would be customary to discuss standard error of measurement. And I would if it were not for the common misunderstanding of standard error of measurement (identified by Dudek, 1979). So how about I just list the equation and move on to the far more useful statistic that actually does the job that ill-informed people (a group that formerly included me) think standard error of measurement does?

Here it is, the standard error of measurement (SEM).

$$SEM = S_X\sqrt{1 - r_{XX}}$$

So, it is pretty simple. It's just the standard deviation times the square root of one minus the reliability. It's a shame that it's not as useful as it seems. (The one bona fide use for the standard error of measurement involves setting cutoff, or passing, scores for tests. More information on how the SEM is used for this purpose is given in its glossary entry.) Let us never speak of it again.

The Standard Error of Estimation

The standard error of estimation (SEE) actually does the job that most people think SEM does (again, see Dudek, 1979). And what is that job? We'll use an example to illustrate.

Let's say a person has an observed score of 80. As we know, odds are extremely high that this score includes some random error. Maybe it contains two points of random error in the positive direction (which raised the score to 80). Or maybe it contains negative five points of random error (which lowered the score to 80). It is unfortunate that we don't know specific values for specific people. But, if we know the reliability of the test, we can estimate the expected magnitude, in terms of points on the test, of random error. Yes, some test takers will have more or less random error in their observed scores than others, but this will be the average number of points. We can then take this estimate and use it to place a confidence interval around the observed score. The confidence interval tells us the likely location of the person's true score, which, as you know, is the part of the observed score that is free from random error.

You have heard the term *standard error* before. The reason you've heard it is that there are many different kinds of standard errors. There's a standard error of the mean, a standard error of a correlation, and the not so useful standard error of measurement. This is the standard error of estimation, a name given by Lord and Novick (1968).

$$SEE = S_X\sqrt{r_{XX}(1 - r_{XX})}$$

So, it is pretty simple. It's just the standard deviation times the square root of the reliability times one minus the reliability. The standard error of estimation tells us the expected magnitude, in terms of points on the test, of random error. By inserting the SEE into another equation, we can find out the likely location of the person's true score. We can never know the exact location, but we can be

95% sure that it is in a certain range. To compute a confidence interval giving the likely (95% is the traditional standard) value of a person's true score, we take the SEE, multiply it by 1.96, and add and subtract it to the person's modified observed score. (Side note: 1.96 is what makes this a 95% confidence interval. In a normal distribution 95% of the scores are within 1.96 standard deviations from the mean.)

To summarize, we know a test taker's observed score. This observed score is affected by random error. We would like to know the test taker's true score, a score that doesn't include random error. Although we can't determine the test taker's actual true score, we can compute a range of values (the confidence interval) that likely includes the true score. The only information required to do this is the test taker's observed score, the reliability of the test (an index of the amount of random error on the test), and the standard deviation of the test.

One might think that this confidence interval equals the observed score plus or minus 1.96 times the standard error of estimation. That person would be wrong. As mentioned before, this confidence requires a modified version of the observed score. Modified in what way? The technical term for this modification is that it is being given a regression to the mean (RTM) adjustment. Don't let the terminology scare you; it's a simple adjustment requiring only the reliability and mean of the test.

$$X_{RTM} = \bar{X} + r_{XX}(X - \bar{X})$$

Now we have everything we need to form a confidence interval giving a range of scores that likely includes the true score for a given test taker. The equation for the 95% interval is as follows.

$$CI_{95} = X_{RTM} \pm 1.96(SEE)$$

Where:
CI_{95} is the 95% confidence interval.

This tells us the likely (to 95% certainty) location of the true score. We can never be 100% confident. If you want to be 99% confident, multiply the SEE by 2.58 instead of 1.96.

Let's run through an example. Our test taker scored an 80 (that's the observed score). The standard deviation of the test is 10, the mean is 60, and the reliability has been estimated to be .90. We'll start with the the standard error of estimation. Plugging into the standard error of estimation equation, we obtain a SEE of 3.0 ($SEE = 10\sqrt{.90(1-.90)}$). Next, we need to compute the RTM adjusted observed score. Given our test taker's observed score of 80, the test mean of 60, and the reliability of .90, the RTM adjusted observed score equals 78 ($X_{RTM} = 60 + .90(80 - 60)$). To finish the confidence interval, we plug the SEE and RTM adjusted observed score into the confidence interval equation and obtain a high value of 83.88 (Upper CI = 78 + 1.96 × 3) and a low value of 72.12 (Lower CI = 78 - 1.96 × 3). We can say with 95% confidence that this person's true score (T) is between 72.12 and 83.88.

Notice that we did not say we are 95% confident that this person's observed score (X) is between 72.12 and 83.88. We know this person's observed score. It's 80. We can say that with 100% confidence (for this test administration). We want to know the value of the true score because the true score doesn't contain random error. Just as with standard errors, there are many kinds of confidence intervals (for mean, variance, correlations, etc). This confidence interval allows us to identify the likely location of the true score for a given test taker.

Just for fun, let's do one more example. We'll keep most of the numbers the same ($X = 80$, $S_X = 10$, $\bar{X} = 60$), but the reliability will be a perfect 1.0. In this case, the SEE is 0.0 ($SEE = 10\sqrt{1(1-1.0)}$), the RTM adjusted observed score is 80, and the 95% confidence interval is 80 +/- 0, which means

that we can be 95% sure that this person's true score is between 80 and 80 (which really means that we can be 100% sure that the true score is 80). Why? The reliability of the test is perfect ($r_{XX} = 1.0$). Thus, this test is completely free from random error. Which means that the observed score is completely unaffected by random error. And because $X = T + e$, the observed score equals the true score ($X = T + 0$). Fascinating, no?

The Standard Error of the Difference

This is very similar to the standard error of estimation. Here, the issue is whether two people's scores are so far apart from each other that random error can be ruled out as the cause of the difference. For example, if my score is 89 and yours is 90, what are the chances that your true score is really higher than mine? (Note that we can say with complete confidence that your observed score is higher because 90 is indeed greater than observed score of 89. But who cares about that?) Stated another way, is your observed score higher because of random error? Unless the test has near perfect reliability or a very small standard deviation, we can't state with any confidence that your true score is higher than mine. This one point difference could easily be due to random error. New example. What if your score is a 90 and my score is a 50? Can we say with any confidence that your true score is greater than mine? Again, it will depend on the reliability and standard deviation of the test. The concept is called the standard error of the difference (SED). The good news is that the SED equation is very simple. The bad news is that until recently it was designed using the wrong key component (see Gaperson, Bowler, Wuensch, & Bowler, 2013). The general form of the equation for standard error of the difference is listed below.

$$SED_{95} = 1.96(SE?)\sqrt{2}$$

Where:

SED_{95} is standard error of the difference (95% confidence)

SE? is the standard error of something.

Let's cut to the source of the confusion, the standard error term that I called *SE*?. Psychometrics textbooks of all kinds will tell you to use the standard error of measurement (SEM). That's right, the standard error that earlier we stated was borderline useless. Our friends Gasperson et al. (2013) did us a favor by extending Dudek's (1979) arguments about the appropriateness of SEE over SEM to the standard error of the difference equation. The short version is that in order to determine the minimum difference between two observed scores that allows us to conclude that the respective true scores are likely different, we should be computing SED using the standard error of estimate. In short, the equation should look like this:

$$SED_{95} = 1.96(SEE)\sqrt{2}$$

Where:

SEE is the standard error of estimation.

So that's it. Just multiply the SEE by 1.96 and the square root of two for 95% confidence. If the difference between the two scores exceeds the SED, then we can be 95% confident that the person with the higher observed score has a higher true score than the other person. Stated differently, we can state that it is unlikely that random error is the source of the difference between the two observed scores. As an example, let's say you score a 90 and I score an 80 ($r_{XX} = .90$ and $S_X = 10$). The SED is 8.32. Because my observed score is 10 points lower than yours and this 10 point difference is greater than the SED, we can state (with the usual 95% confidence) that your true score is greater than mine. (The technically correct version of the preceding statement is: The 95% confidence interval for the difference between our scores is less than the observed difference.) Conversely, we

could state that it is unlikely your observed score is greater than mine simply because of random error.

Let's do one last example where everything is the same (my score is 80, your score is 90, $S_X = 10$), but the reliability is perfect ($r_{XX} = 1.0$). As we saw earlier, the SEE equals zero, which means the SED equals zero. Which means that we can be completely confident that your true score is greater than mine. In fact, even if my observed score was an 89 (only one point behind yours), we could still be completely confident that your true score was greater. Why? Perfectly reliable test. No random error of measurement. You get the idea. But don't forget: True scores include systematic errors. Maybe that's why you did better. (Gotta protect my ego somehow.)

Item Writing and Test Construction

"I have made this letter longer than usual, only because I have not had time to make it shorter."

-Blaise Pascal

Introductory Remarks

How are tests made? The answer to that question is a long and interesting story. And here it is.

As mentioned in Chapter 1, we measure behaviors on a test in order to infer a person's standing on a construct. If we are trying to identify who is smart, we are measuring the construct of intelligence. If we observe a person answering a large number of difficult math and verbal questions correctly (the responses to the questions are the behaviors), we infer that this person is smart (i.e., has a high standing on the construct of intelligence). Second, tests are composed of one or more items. What's an item? In short, an item is a synonym for a question. For more detail on this, see Chapter 1.

Can we have a one-item test? Yes, we can, but we probably don't want to. The reasons are threefold. First, from the test-taker's perspective, test takers would not like a one-item test as the chance of scoring a zero is too big to be palatable. Suppose I offer you a chance to take one item, chosen randomly, from my 50 item final exam. If you answer it correctly, you get a 100. Miss it, and you get a zero. Would you do it? I've made that offer (and have been in classes where that offer was made), and I have never seen anyone take it. Because even if you study thoroughly for the test and know 95% of the material, if the one question is from the 5% you don't know, you could wind up with a gigantic zero staring you in the face for the final exam grade. If you recall from the first chapter, a test is a sample of behavior. Based on this sample of behavior, we infer your standing on the construct. If the test has only one item, that is a poor sample of behavior. Any inference based on this single behavior is likely to be wrong.

This issue concerns the content validity (also called domain sampling adequacy) of the test. The purpose of some tests is to generalize performance from the test (e.g., the road part of a driver's test)

to something larger than the test (actually driving a car). The written driver's licensing test is another example, in this case, it's knowledge based. We need to know if a person knows what to do at a four way intersection, a three way intersection, when you need to use the turn signal, and so on. Now imagine that you take the written test and there is only one question: What does a flashing yellow light mean? Your thought probably would be, "What about all of that other stuff? They never even asked me about any of the issues related to right of way. How odd. I should alert my measurement professor." After a while, it might occur to you that there will be a whole bunch of drivers on the road who don't have a clue about how they are supposed to drive. All because the state used a one-item test for licensing. Getting back to this content validity thing. It should be obvious that a one-item test will have a hard time asking about all of the important issues that we want people to know. And their performance on this one item will not be representative of their standing on the construct.

There is another reason for avoiding one-item tests. Namely, the role of random error is too big. In a one-item test, a person who guesses correctly (and thus deserves a zero), gets a 100. That means that 100% of his or her observed score is due to random error. Similar problems apply to the person who actually knew the correct answer and accidentally wrote down the incorrect answer. Ouch! And this whole time you've been thinking that the one question was a well-written one. Imagine if it happens to be a poorly written, hard to understand question. Not good. Now if we have two items instead of one, there is at least an opportunity for a random error in one item to offset a random error on the second item (i.e., I miss Item 1 when I actually knew the answer, but guess correctly on Item 2 when I didn't know the answer – thus, my overall score ends up being correct: 50%). This may sound like a long shot, and it is,

but at least that possibility is there. With three items, I have even more opportunities for random error to cancel out (and of course, the magnitude of a single random error is reduced). With four items, even more, and so on. This is an illustration of the principle behind the Spearman-Brown prophesy formula (see Chapter 5). Given items of equal quality, longer tests are more reliable.

So one-item tests are to be avoided. Moving on, another point from Chapter 1 that bears repeating is that we can measure multiple constructs on a given test. Now I'm quoting here, "Important point: We're used to thinking of a test as what we can fit on a few sheets of paper – one test booklet, one test. But the number of distinct tests (as indicated by the number of constructs measured) is determined by the number of ways we *score* the questions on the paper. One test booklet may contain as many tests as you like" (myself, this book, somewhere in Chapter 1). Also, as mentioned in Chapter 1, we typically, but not always, want our tests to be unidimensional. That is, each test measures a single construct. If I want to measure five constructs, I make five tests. I may even put them on the same piece of paper, but they are still five unidimensional tests.

Now that we have that out of the way, I can tell you how we make tests. It's a ten step process. Why ten steps? It's a nice, round number.

Step 1: Why? Define the Purpose of Testing

Why are we testing? Is it for licensing purposes or because we want to separate and identify the individuals with high standings on the construct from those with average and low standings? This is a big issue and determines what type of item analysis (see Chapter 9 for item analysis) we favor. Who is the target population? If our test will be given to a low ability population, then we need to write easy questions. What about reading level? Unless our test is a measure of reading abil-

ity, we need to consider the reading level of our target population and make sure that all written instructions, prompts, and questions are written at a level that is safely below that of our target audience (remember reading level describes the performance of an average person at that level – half of the people at that age or grade will not be reading at that level). A popular tool in industrial-organizational psychology was unintentionally written at a reading level far in excess of its target audience. As a result, the test cannot be directly given to employees for them to answer on their own. Rather, a consultant has to interview the incumbent and answer the questions for him or her. A better strategy would have been to rewrite the test at an appropriate reading level, but for reasons known only to them the test developers went their own way.

Step 2: What? Identify What Will Be Measured

We'll move from the construct to the behaviors. Remember that constructs are assumed to cause our behaviors. And on a test, the behavior is the response to the test item. So, the main question is: What construct will we measure? Once you have chosen your construct, it's time to narrow matters. Some constructs are broad and some are specific. If we want to measure intelligence, that's a pretty broad construct. There are many aspects to intelligence. Which ones do you want to measure on your test? All of them? Just the main two or three? Just the ones for which there is general agreement in the field? What we are doing is identifying subconstructs, which are nothing more than more specific subcategories of our main construct. If our construct is intelligence, then our subconstructs might be math, verbal, spatial, and memory. Thus, it helps to define these subconstructs. Now we have a much better idea of what we want in our test and what we want out of it. Of

course, some constructs are already so narrow (like attitudes about Coke versus Pepsi) that identification of subconstructs isn't possible.

The next step is to further narrow matters and define the content domain. The content domain is the set of all behaviors and/or information relevant to the purpose of the test. And for most tests, the purpose is to measure an unobservable construct. (There are also tests that are designed to simply measure a set of behavior with no reference to unobservable causes. For more information, see the discussion of the definition of construct in Chapter 1.) It is here that for most tests we make the leap from the unobservable (the unobservable construct) to the observable (behaviors). Imagine that we are trying to measure depression. The content domain consists of all possible behaviors related to depression. Things like a lack of energy, a change in sleeping and eating habits, and so on. Obviously, listing all possible behaviors is going to be a big job. In some cases (e.g., math) it will be an impossible job. But we need to do it. For something like depression, we can consult professional manuals (e.g., the DSM) for help. For math, we can never make a list of all possible relevant information due to the nature of the number line, which I hear goes on forever, though I haven't checked. We can however, define limits to our domain. That is, what is in and what is out. We might decide that addition, subtraction, multiplication, and division are in, but calculus is out. Geometry is in, but trigonometry is out. Bear in mind that ideally, we would list all of the relevant behaviors and/or information. We'll settle for identifying the limits when a complete listing is impossible. One final example of defining the content domain: What is the content domain for a test in a social psychology class? It would be every piece of material that was covered in the lectures plus everything assigned in the book and/or other materials. Thus, it is fairly easy to define the content domain for some tests (e.g., a test in a social psychol-

ogy class) and more difficult for other tests (e.g., math).

Once we define the content domain, we have to realize that we won't be able to have a test question for everything in the domain. Even for a test from a relatively limited content domain (like a test in a social psychology class), we can't ask about everything, or we would have a test around a hundred questions, an impractical length. Now consider a broader construct like depression or math, and you've got a test with hundreds, or thousands, of items. Not practical. Thus, our tests must *sample* from the content domain. We want the items on our tests to be a representative sample from the behaviors/information in our content domain.

The final issue is to introduce a tool to help us write items. Item production rules (IPRs) are simply rules that we make to guide the writing of our test items. An example of an IPR for a math item would be: A three digit number divided by a two digit number for which there is not a remainder. Consider how this IPR would help us. We can come up with dozens of items that fit this rule very easily. Everyone can understand the IPR and the items that they generate will be roughly equal in difficulty. Very helpful. But that wouldn't be the only IPR for our test. We would have other IPRs to help us write other items. Another IPR might be: A three digit number divided by a two digit number for which there is a remainder. A slightly different kind of item with a different difficulty, but still easy to generate. Naturally, our math test would need plenty of IPRs to cover all of the different items in our content domain.

The math test IPRs I just cited were fairly specific. Other IPRs will be less structured. Consider a test of depression. Our IPRs won't be content based, rather they will indicate the general structure of the item. Here is an example: Each item will consist of a behavioral statement (not a ques-

tion) followed by a three-point response scale indicating agreement in some way. Now consider the following two items.

1. Do you feel sad?

   A. No

   B. Maybe

   C. Yes

2. I often lack energy.

   A. Disagree

   B. Neutral

   C. Agree

It should be clear that the first item does not fit our IPR (it's a question, not a statement) whereas the second one does. It is unfortunate that our IPR can't cover the content of the construct like our math IPR, but that is just the nature of the construct. The content domain we defined earlier will help supply the content of the items.

Step 3. How? Which Items Will Be on the Test?

We need to write a plan for our test. Really, it's like an outline. How many of each type of item will we have on our test? These plans are called test specifications. Following are some examples of test specifications.

I. The test will consist of 50% multiple choice and 50% short answer.
II. The test will have 25% addition, 25% subtraction, 25% multiplication, and 25% division questions.
III. The test will consist of 38 vocabulary questions and 12 reading comprehension questions.

In every case, the above test specifications help us plan our test. The first test spec is a structure one, it tells about the format of the items, whereas the last two are content specs, they refer to certain parts of the content domain.

What are test specs good for? They tell us how many items we need to write of each kind. They help us determine if we have sampled appropriately from the content domain. They tell us if we need to write more items of a given type in order to sample appropriately from the content domain. And finally, if given to test takers, they help test takers study for a given test by directing their effort to the appropriate areas.

Step 4. Write Items

Finally, we get to put pen to paper and write some items. This process will be a world easier if we have (a) defined the content domain (or at least the boundaries to the content domain), (b) written IPRs, and (c) written test specifications. Writing items is, in large part, a creative process. This is where good, original ideas can pay off in a big way.

Step 5. Edit Items

Once written, the items, like any project, need to be edited. We should review our items to make sure that they are: (a) accurate, (b) clear, (c) have correct grammar and punctuation, (d) written at the appropriate reading level, (e) maintain a consistent tense, and (f) minimize negatives (e.g., "Which of the following is not true."). Additionally, optimal performance items should have an even allocation of correct (or *keyed*) answers across alternative positions. Some computer programs that assist with test construction will randomize the order of the alternative answers. Inventory type tests should have the stems written in a mix of both positive (e.g., "I like candy.") and negatively oriented statements ("Candy makes me sick.").

A few things to avoid in item writing are: (a) inequities in alternative length, (b) double negatives, (c) universals (e.g., always, never), (d) syn-

onymous or implausible distracters (distracters are the wrong answers in optimal performance multiple choice items). Finally, one should avoid measuring two constructs on the same item (referred to as a double-barreled item). If you want to measure two constructs, write two different items (or better yet, two different tests).

Step 6. Item Tryouts

At this point we have a set of items that we think are written well. It's time to get an outsider's opinion. In fact, we don't want just any outsider, we want the opinions of the same type of people who will be taking this test. If our test is designed to be given to high school seniors, we want to get a small group of high school seniors to review the items. (It should be obvious there are some ages for which this plan will not work; in those cases we can get opinions from those close in age.) We want our item tryout group to take the test and tell us any problems they see with item wording or design. We are most interested in knowing if they interpret the items in the way in which we intended. One way to conduct item tryouts is to give our test takers a tape recorder and have them verbalize their thoughts as they take the test. As should be obvious, we will not be able to handle a large group of people for our item tryouts as we will have a great deal of data to examine for each person. One thing that we do not care about are the scores of the people in our item tryout sample. That's a separate issue and will be discussed next. Here, our major concern is whether the items were written appropriately.

Step 7. Item Analysis

Item analysis is the phase of test development in which we collect actual data from test takers, analyze the data, and decide which items to keep in our test. Unless you like having bad items on your test, you will not keep all of the items. Thus, you should enter the item analysis phase with

more items than you will ultimately need. Item analysis will be discussed in detail in Chapter 9. So that's something you may look forward to (to which you may look forward?). It is important to note the difference between item tryouts and item analysis. With item tryouts, we want the qualitative opinions of the test takers regarding their understanding of the items. Item tryouts are judgmental steps that usually involve editing the items. In item analysis, we examine the answers that test takers give to the questions (i.e., the actual quantitative data). It's a statistical analysis which usually leads to our deleting bad items.

Step 8. Reliability Analysis

We've already covered how we estimate reliability with classical test theory. You will recall that the reliability of a test indicates the test's freedom from random error. If you do not recall this, you might want to give Chapter 4 another run. The importance of finding out whether our test is riddled with random error should be obvious to all. If you want to estimate reliability with the alternate forms method, go back to Step 4 and write a second version of the test. Other than that, just follow the instructions in Chapter 5 for estimating reliability. One important concern is that our item analysis and our reliability analysis should be conducted with separate samples of data. If not, we will likely obtain a biased estimate of reliability (particularly if we perform an internal consistency based item analysis followed by an internal consistency estimate of reliability).

Step 9. Validity Analysis

In addition to simply being reliable, a test should be valid for the purpose for which it is used. Thus, we need to establish the validity of the test. This is another long topic which will be discussed later (Chapter 10). Do we need a new sample of data yet again? No, if you're talking about the reliability and validity studies – they can

share the same sample of data. Yes, in that our validity sample needs to be different from out item analysis sample (particularly if we conducted an empirical keying item analysis). A big picture issue to keep in mind is that item analysis is an analysis of individual test items whereas reliability and validity studies are about the test as a whole. The test is in its finished form when we start our reliability and validity analysis.

Step 10. Write Test Manual, Administration Guidelines, and Scoring Procedures

At this point, we have a test that is in its final form and is useable. Thus, the research stage is over and it's time to put the test in operational use. That's just a fancy way of saying we can use the test. Depending on the type of test, the labor of a scientist is no longer needed. Many tests can be administered and scored by a clerical worker. Or even a ten-year old. For this to happen, we need to take the final step and write administration and scoring guidelines. Furthermore, we need to write a report of our test development process, with a heavy emphasis on the reliability and validity results.

Concluding Thoughts on the Steps

And that's that. As an aside, because we ultimately care more about validity than reliability, one could skip the reliability study and go straight to the validity study. However, given how easy it is to compute a coefficient alpha estimate of internal consistency reliability, there's not much reason to skip it. Moreover, let's say your test is riddled with random error and you don't know it because you got lazy and skipped the reliability step, you could be unknowingly wasting your time with the validity study because an unreliable test (once again, unreliable means random error) can never be a valid measure of anything (except randomness, I suppose).

Challenges in Item Writing

One factor which makes it difficult to write good items are time and space constraints – a test has to be completed in a limited amount of time, and the test item must be as brief as possible. This is a challenge because given unlimited space we could clarify ourselves so that the test taker understands what it is that we are asking. However, in a limited amount of space we don't have the luxury of explaining what we do and do not mean in detail. We have to clearly communicate an idea in a limited number of words. That is a serious challenge (see the quote at the beginning of this chapter).

The result of these space constraints is that we often write items that make perfect sense to us, the item writer, but are confusing to the test taker. We weren't trying to make them confusing, but because we couldn't explain ourselves fully, ambiguity was introduced. (Side note: I once had a professor attempt to clarify his test question by reading it aloud to the class with a pause at a certain spot in the sentence. He thought this was helpful. I did not. And yes, I remember this only because I ended up missing the question. At least I got a good story out of it.)

Given this challenge, we must work even more diligently on our writing and editing to minimize the number of times that we end up with ambiguous items. We must also make use of item tryout information to identify and revise these poorly written items. It's work, but it must be done.

Further Thoughts on Item Difficulty in the Item Writing Process

When a test taker answers a right/wrong item, two outcomes can occur: a correct or incorrect response. Yeah, that's obvious. But we must consider whether they deserved to make the correct or incorrect response. We understand that ran-

dom errors influence this result. Sometimes you guess correctly when you didn't know, and sometimes you miss something that you did know simply because you wrote "A" when you meant "B".

Poorly written items also result in these two errors. Items that include too much information (i.e., they give away too much) or have poorly written distractors help test takers answer correctly when they should have missed it. Items that trick test takers out of making correct responses cause them to miss those items when they should have answered them correctly. Writing intentionally confusing items is no badge of honor for test developers – they have purposefully created measurement errors. It should be seen as a mark of shame. Giving away the answer and confusing test takers so that they make the wrong answer are both bad outcomes. Chart 1 illustrates the various scenarios that can occur when a test taker answers an item.

CHART 1 Possible Measurement Outcomes

	Test Taker Does Not Have Enough Ability	**Test Taker Has Enough Ability**
Test Taker's Response is Incorrect	Test taker should have answered incorrectly and did (this is a desirable scenario)	Test taker should have answered correctly but missed it; trick (or ambiguous) question (not desired)
Test Taker's Response is Correct	Test taker should have missed it but answered correctly; item gave away helpful information (not desired)	Test taker should have answered correctly and did (this is also a desirable scenario)

You don't have to look at the chart for long to think, "Isn't this just common sense?" The answer is yes. But these rather obvious issues should always be in mind during the item writing process. Consider this the fundamental philosophy of item writing: write items that will only be answered correctly by those who have sufficient ability or knowledge and will only be missed by those without sufficient ability or knowledge. Given the existence of random errors, we know that this outcome will be impossible to achieve, but it is the goal to which we aspire.

Item Types

7

Limitless options.

Infinite possibilities.

And two very popular item types.

Introduction

This chapter will illustrate several common item types, though more exist than we'll cover. We'll spend a little extra time on Likert items because they are used extremely frequently.

Optimal Performance Items

We'll just mention the basic multiple choice and true/false designs. One thing to note about these kinds of items is that they are almost always scored in a dichotomous fashion. That is, there are only two possible scores, usually 0 and 1. A multiple choice item may have five alternatives (A through E). But once it is scored, there are only two numbers. If C is the correct answer, then a C response gets a 1 and all other responses get a 0. Most of the other issues were covered in Chapter 6, so we'll move on to other item types.

Ranking. What follows is an example of a ranking item.

> Rank the importance of the following objects to your survival on a camping trip gone awry.
>
> _____ Canteen
>
> _____ Map
>
> _____ Compass
>
> _____ Cellular Phone
>
> _____ Landshark Repellent Spray
>
> _____ Bag of Donuts

Our job is simple. We'll use the numbers 1 to *n* to describe how important each of these items are to our survival. Scoring this test will require a key. How we obtain the answer key, that is, the correct ranking, is another issue, but once we have the key, we compare the obtained numbers to the numbers on the key. Let's say the answer key says the Canteen of Water is "2". If we answered "3", we

would be a point off and thus, receive a point. Now let's say the keyed answer for Map is 1 and we put 1. No points here. It's a perfect match. We repeat and sum up to get a total score. Just like golf, the lowest score wins. Closely related to ranking is point allocation.

Point Allocation. Here's an example of a point allocation item.

> Describe how important the following objects are to your survival on a camping trip gone awry. You have a total of 100 points to assign. More points indicate greater importance. You may assign points in any way you like as long as the sum of the points equals 100.
>
> _____ Canteen
>
> _____ Map
>
> _____ Compass
>
> _____ Cellular Phone
>
> _____ Landshark Repellent Spray
>
> _____ Bag of Donuts

How do these two methods differ? With ranking, we can only show relative differences in importance. I might say that the Map was more important than the Canteen of Water, but I couldn't show just how much more important I considered it. This is the old ordinal versus interval level of measurement issue. With point allocation, I can give the Map 90 points, the Canteen 10 points, and everything else 0 points. Now you can see that I strongly value the Map. Scoring is just like before, the closer my point totals match the keyed answers, the better my score. Here's something to think about: Is point allocation interval or ratio level measurement?

If a test taker wants to get cute, they might do something like this:

20	Canteen
30	Map
-10	Compass
50	Cellular Phone
0	Landshark Repellent Spray
10	Bag of Donuts

This is a problem. It would be a good idea to add a "no negative numbers" clause to the instructions. Besides, if zero means not important at all, then there really is no way to have an option for something less than that.

One final thought on point allocation. What if the list of items is long, and our test taker isn't the biggest fan of math?

Items for Measures of Attitude and Personality

Forced Choice. Which of the following describes you better?

Talkative	Shy
A	B

This item is really a slight variation of a true/false question applied to attitudes/personality. You get a point for picking the keyed response, and you get nothing for any other response. No allowance is made for in-between responses.

Semantic Differential. Which of the following describes you better?

Talkative				Shy
A	B	C	D	E

With a semantic differential item, we've taken a forced choice item and stretched it out to a five point scale. If I feel completely neutral on the

item, I can pick C. If I feel that I am a little shy, I can pick D. And so on. Scoring is the same as Likert items, which are next, so we'll just move on to them.

Likert. With these items, we will make a statement or ask a question and offer a multipoint response scale. The two most common response scale types (frequency and agreement) are offered below. An infinite variety of other types are possible. There's this kind, in the form of a question.

How often do you hear voices?

1. Never
2. Sometimes
3. Often
4. Always

And this kind, in the form of a statement.

I like to hug total strangers.

1. Strongly Agree
2. Agree
3. Neither Agree Nor Disagree
4. Disagree
5. Strongly Agree

Scoring is simple. Pick a keyed direction for each item. Let's take our "hugging" item. Let's say that we want high scores to mean that people like hugging. As it stands now, high scores mean that people don't like hugging. We need to *reverse code* this item. To reverse code, all we need to do is invert the coding. We'll change the numbers so that a 1 will now be a 5, 2 will now be a 4, 3 is still a 3, 4 is a 2, and 5 becomes a 1. Once we've done this, the scores on this item are now consistent with our desired interpretation: Higher scores indicate greater fondness for hugging strangers. That's it. Notice that we do not have dichotomous scoring (only two possible values) here. We have multi-

level scoring in which there are more than two possible values. In this case there are five.

Another issue with Likert items concerns the anchors. How many anchors should we have? Should we have a middle anchor, allowing a neutral opinion? There has been much research over those issues and as long as there are not more than nine anchors, nothing much matters as regards those two issues. An issue that is of primary importance is what we use for our anchors. Can we leave anchors unlabeled (also called unanchored points)? We can, but it is a bad idea. To understand why, let's discuss why we have anchors. Imagine that you see this item.

I like to hug total strangers.

1 2 3 4 5

I'm the test taker, and I don't like to hug strangers at all. How do I make my response to indicate my attitude? They are not giving me a lot of information to work with here. Maybe bigger numbers mean I like hugging strangers more. Maybe not. How do I pick a response to indicate my attitude? I understand the question and I know how I feel about it, but how do I respond to indicate my position?

It should be clear that we use anchors to help people make a response that indicates what they are thinking, feeling, or do. Now look at this item.

I like to hug total strangers.

1 2 3 4 5 6 7

Agree Disagree

If I am the test taker, I now have more information to use when making my response. Greater numbers mean I dislike hugging strangers. But what if I only sort of dislike hugging strangers? Do I pick 4, 5, or 6. Maybe not 4 as it appears to be a neutral position. (Why didn't they just label it as "neutral"?) So, will it be 5 or 6?

Again, our test taker doesn't have enough information. Unanchored scale points create ambiguity and allow for more error and bias in the responses. And the last thing we need is more opportunity for error and bias. (As an aside, this issue speaks against semantic differential items, although the process there is slightly different.)

Finally, it is not enough to label the anchors with just anything, we need to have clear, understandable anchors. Look at this item.

How often do you hear voices?

1. Never
2. Seldom
3. Occasionally
4. Often
5. Constantly

Now look at this one.

How often do you hear voices?

1. Never
2. Occasionally
3. Seldom
4. Often
5. Constantly

Does one look obviously right and the other obviously wrong? I switched the second and third options. If those were clearly defined, well chosen options, then one should look obviously wrong. The fact that neither one does is a sign that the anchors are not clearly defined. We want a clear difference between anchors from two different scale points, and we don't have it. What follows is an improved (but by no means perfect) version of the same item.

How often do you hear voices?

1. Never
2. Occasionally
3. About Half of the Time
4. Often
5. Constantly

Still a better version would be:

How often do you hear voices?

1. Never
2. No more than once a year
3. No more than once a month
4. No more than once a week
5. Many times each week

Closing Thoughts

And that's just the tip of the iceberg when it comes to the different types of items that exist in standardized testing. Sure, the classic multiple choice and the standard Likert item are the most popular formats, but the possibilities are limitless.

Composite Variables

8

Throw a rock over your shoulder, and you'll hit a composite variable.

Or two.

Introduction

We discussed one-item tests before. Given the problems with one-item tests, it's clear that we want tests with multiple items. Here's something to consider: How do you report the scores on a multi-item test? Let's say someone takes a ten question spelling test. Do we report his score as ten individual scored responses (e.g., 1, 1, 0, 1, 0, 0, 1, 1, 0, 1)? Of course not. We form a total score, a mean score, or some other combination of the individual item scores (using the data from the previous example, we would report a 6, or 60%, or something like that). Any combination of scores on individual items into some number is a composite variable (or composite score).

Composite variables have some useful properties, some of which we have already discussed. One property was mentioned in the above paragraph; that is, composite variables are convenient summaries of performance on multi-item tests. Another desirable property was covered in Chapter 5. In the explanation of internal consistency reliability in that chapter, one of the things we saw was that, other things being equal, longer tests are more reliable. How does this relate to composite variables? It should be clear that a total score across ten items is more reliable than ten single-item scores (assuming test items of equal quality). So, the long and short of it is we like composite scores. We use composite scores almost all of the time. Let's dig into two important characteristics of composite variables.

Mean of a Composite Variable

This one is fairly obvious. How do you think individual item means relate to the mean of a composite (let's say a simple total score) of these items? If I told you that I had a 10-item test and each of these ten items was answered correctly only twenty percent of the time, what do you think each test taker's total score would be? High?

Medium? Low? Obviously, low. And of course the mean of these total scores would be low. You knew that. It's obvious. If all of the test takers are missing the test questions like crazy, then there is no way for the mean of the total scores of these test takers to be anything other than low.

New example. It's a 10-item test, and every item is answered correctly by every person. What do you think the mean of the total test score will be? Let's see, every person answered all ten questions correctly. It sounds like each person will have a total score of 10. And if each person has a total score of 10, then the mean of these composite scores will be 10. There's no way for it to be anything else.

What can we conclude from our little exercise? We conclude that the mean of a composite variable is directly linked to the means of the items. If the composite variable happens to be simple total score, then we can summarize the rule in this handy fashion: The mean of a sum equals the sum of the item means.

You want an example? OK, here's an example. Each item in the example below is scored with one point for a correct answer and zero points for any other answer. The composite score is the simple sum of the correct responses. The last row is the percent of people who answered the item correctly.

Person	Item 1	Item 2	Item 3	Item 4	Total
Mary	1	0	1	1	3
Beth	1	1	1	0	3
Billy	1	0	1	0	2
Mr. Mentalino	0	1	0	0	1
Harry	0	0	0	0	0
Percent Correct	0.6	0.4	0.6	0.2	-

Now that we have our data, let's check the rule we mentioned (the mean of a sum equals the sum of the means). The percent correct for a dichotomously scored item is the item mean. (If you're not sure, let's look at the Item 1 data to check this. The sum of the scores divided by the number of scores equals .6.) The item means are .6, .4, .6, .2. These four values sum to 1.8. So that's the *sum of the means* part. What about the *mean of the sum* part? The total scores are 3, 3, 2, 1, and 0. The mean of these five scores is 1.8, the very same 1.8 we obtained before. Why did we go to all of this trouble? To illustrate that the mean of a composite variable is directly linked to the mean responses to the individual variables that form that composite.

Variance of a Composite Variable

The second important characteristic of composite variables is variance. Let's think back to some concepts from the first few chapters. First, variance is about differences in scores. A set of scores with a variance of zero means that everyone has the same scores. Second, variance is good. We want the scores to be different. Why? Because we assume that different people have different standings on the construct, and the test should accurately reflect these differences. A test that gives everyone the same scores has failed in its job. (Either that, or everyone actually has the same standing on the construct, an unlikely scenario.) Also, we hate tie scores. Third, given a constant level of measurement precision, bigger differences between scores allow us to be more confident about the differences between the scores.

What determines composite score variance? It's not as simple as what we saw with the mean of a composite variable. Part of the equation is item variance. No big surprise there. Other things being equal, more item variance means more composite score variance (aside from a very weird exceptional case). What kind of items have a lot of

variance? Items that are average in difficulty. Why? An item answered correctly by everyone has no variance (all scores are the same; e.g., 1, 1, 1, 1). The same applies to an item that everyone misses (e.g., 0, 0, 0, 0) – no variance. It should be clear that item variance is maximized when half of the test takers miss it and half answer it correctly. But there's more to composite variance than just the item variances. What else? Something like the correlations between the items.

The exact statistic involved is covariance. Remember covariance from some earlier chapter? I don't either. I found this equation. Maybe this will help.

$$c_{XY} = r_{XY} \cdot S_X \cdot S_Y$$

So it appears that covariance is just correlation multiplied by the standard deviations (and standard deviation is just the square root of variance).

Every item has a variance. Every pair of items has a covariance. Put those together to get the composite score variance equation: The variance of a sum equals the sum of the item variances plus two times the sum of the covariances between each pair of items. That wasn't so bad. The only weird part is that we are supposed to multiply the sum of the item covariances by two. Why do we have to multiply by two? You just do.

(If you're familiar with something called the variance/covariance matrix, the composite score variance equation can be stated as follows: The variance of a sum equals the sum of all of the elements in the variance/covariance matrix.)

Here is an example to show the calculations. In this dataset, we have a two-item test taken by four people. Both items are answered correctly by half of the test takers. The item variances are .33 for each (feel free to check this using the sample variance equation from Chapter 2). The correla-

tion between Items 1 and 2 is 1.0. Using the covariance equation, we find that the covariance is .33.

Person	Item 1	Item 2	Total
Ricky	1	1	2
Blake	1	1	2
Shelley	0	0	0
George	0	0	0

Now to the computation of composite score variance. Variance of a sum equals the sum of the item variances plus two times the sum of the item covariances. In this case, that would be .33 + .33 + 2(.33), which equals 1.33. Can we check this number? Yes, we have the actual total scores for each test taker, let's just compute the variance of these scores (2, 2, 0, 0) to check. A quick calculation of the variance of the total scores shows us that we were correct; the variance is 1.33. So the equation is accurate.

The only remaining issue is, why? Why is the relationship between the items important to composite score variance? Here are two sets of scores: 2, 2, 0, 0 and 1, 1, 2, 0. Which set has greater variability? If we run them through the variance equation, we find that the first set (2, 2, 0, 0) has a variance of 1.33, whereas the second set (1, 1, 2, 0) has a variance of .67. (If you find this outcome a little confusing, here's why the first set has greater variance. Variance is all about how far the scores are from the mean. The mean of both sets is 1.0. In the first set, none of the scores are at the mean; they are all one point away from the mean. In the second set, half of the scores are at the mean, and the other half are one point away from the mean. Thus, there are fewer differences from the mean in the second dataset.)

Back to our two sets of scores. If these are total scores on a two-item test, and if we fix it so that every item is answered correctly by half of the test takers (to make for a fair comparison), how

do you end up with total scores of 2, 2, 0, 0? There's only one way, and it requires our variables to be correlated 1.0. If you would like to see what that data would look like, just check the previous dataset. What about the other set of scores (1, 1, 2, 0)? Given the conditions governing this demonstration, there's only way for a two-item test to yield these total scores. And it looks like this:

Person	Item 1	Item 2	Total
Steven	0	0	0
Lloyd	0	1	1
Rachel	1	1	2
Ron	1	0	1

The correlation between Item 1 and Item 2 is 0.0. Are you starting to see why the correlation among the items matters to composite score variance?

Before I jump ahead to the explanation, one last dataset: 1, 1, 1, 1. What's the variance of these scores? Zero. No variability at all. Every score is at the mean. If, as before, these are total scores from a two items test, with each item answered correctly by half of the test takers, then there's only one way the the data can yield these total scores.

Person	Item 1	Item 2	Total
Elmore	0	1	1
Hubert	1	0	1
Johnnie	0	1	1
Willie	1	0	1

And the correlation is... wait for it... -1.0. What's the lesson? A strong, negative relationship between the items causes the item variance to cancel out when the items are summed to a total score. The composite score actually has less variance than the individual items.

Let's see if we can summarize these observations in a meaningful way. A negative inter-item

correlation causes the total score to have less variance than the individual items. A zero inter-item correlation causes the total score to have the same amount of variance as the individual items. A positive inter-item correlation causes the total score to have more variance than the individual items. Variance is about differences in scores. Positively correlated items allow the differences between people on individual items to accumulate when these items are combined to form a total score.

Final Thoughts

It would appear that the ideal test would be composed of a large number of items average in difficulty and correlated perfectly with each other. Data from such a test would look like this:

Person	Item 1	Item 2	Item 3	Item 4	Total
Alex	1	1	1	1	4
Neil	1	1	1	1	4
Garrett	1	1	1	1	4
Leonard	0	0	0	0	0
Quendra	0	0	0	0	0
Sean	0	0	0	0	0

But is this really an ideal situation? How could you make it happen? Here's a way: Write one item average in difficulty and repeat it four times. Such a test would have the appearance of being desirable due to the high variance in the total score, but it's obvious that we're just kidding ourselves. To use a sports metaphor, it would be playing for the stats and not for the win. Each item on a test is supposed to provide, in part, some unique information. That is clearly not the case when you repeat items. Also, this approach leads to a lot of tied scores.

So we want positively correlated items, but we don't want perfectly correlated items. What about the difficulty part? Do we want all of our items to be exactly average in difficulty? The answer to that question can be found in the next chapter.

Item Analysis

Test day for the test items.

9

Introduction

Item analysis is the process of identifying the good items in our item pool. We do this any time we write new tests (as we learned in Chapter 6). We also do this when we take an existing test and modify it to make it perform better. One of the most widely used personality tests is the MMPI. For over forty years every item was the same as the day it was published. Eventually the MMPI was revised; some poorly-performing items were deleted and replaced by new items. The revised version was given the clever name of MMPI-2.

Let's remember that a test is a collection of items, an item is a single response to a single stimulus, and we would rather not have one-item tests. These issues were addressed in Chapters 1 and 6.

We can accomplish a number of objectives with an item analysis. The two most common objectives are (a) to increase the unidimensionality of our test and (b) to make our test a better predictor of some external variable. Other objectives include the deletion of biased items, adjustment of test difficulty (making the test harder or easier to better fit the test takers' abilities), and establishment of the content validity of the test.

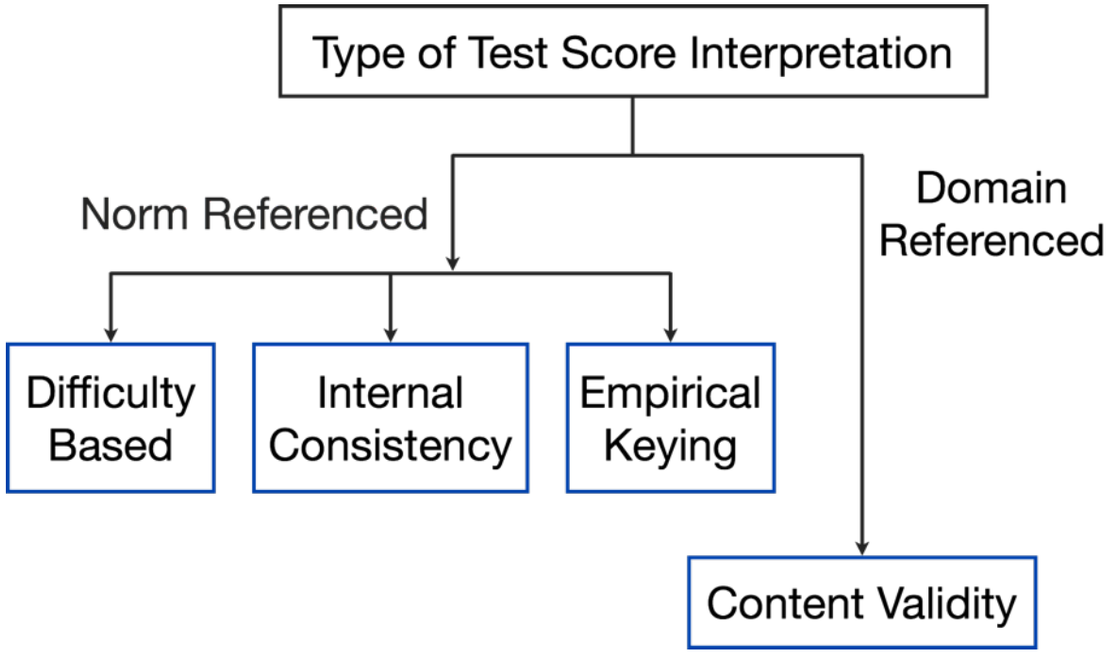
Norm-Referenced Versus Domain-Referenced Testing

We have previously discussed how meaning is given to a test score (see Chapter 2). The most common method is normative inference in which a person's score is compared to the scores from other test takers (we'll give normative inference another name: norm-referenced testing). There is a different method for giving meaning to test scores called domain-referenced testing (also called criterion-referenced testing, but we'll avoid using that name because it's wretched). Thus, you could say that we have norm-referenced testing and domain-referenced testing. With domain-

referenced testing, a person's score is compared to an objective standard, quantified with a cutoff score. An example of this is the licensing exam. With a licensing exam, all that matters is whether a person's score is greater than the cutoff score. If the score is greater than the cutoff score, we infer that the person can perform the job (e.g., driving a car, performing brain surgery, etc.) at an acceptable level. Unlike norm-referenced testing, we don't care how well this person's score compares to everyone else's score. His score may be the lowest score we see that whole month, but if it is greater than the cutoff score, the person passes and we infer that he is qualified. This issue of how we interpret the score (domain-referenced versus norm-referenced) is important because it determines which type of item analysis we perform. Chart 1 shows all of the major item analysis techniques.

A note on the chart. We are not limited to performing just one type of item analysis. We can do

CHART 1 Item Analysis Techniques



two, three, or four different item analyses on one set of items. However, odds are strong that one issue (e.g., unidimensionality, content validity, prediction of external variable) is the primary concern. Enough with the preliminaries, let's get down to business.

Difficulty-Based Item Analysis

The goal of a difficulty-based item analysis is to match the difficulty of the test to the ability of the test taker. That is, if we have low ability test takers, we want our test to be composed of easy items. If we have high ability test takers, our test should be composed of difficult items. And of course, we want items of average difficulty for people of average ability. Why do we want this match? We'll answer that question by exploring another question: What if there is a mismatch of ability and item difficulty? If we give a hard item to a low ability person, two things can happen (the response will either be correct or incorrect), and neither are very informative. If the person misses the item, that doesn't tell us much, because he was of low ability and it was a difficult item – it is what we expected. But what if the person gets it right? That also is not informative, given that we know the item is difficult and the person is low in ability – he likely guessed correctly. It goes the other way too. What if we give an easy item to a high ability test taker? If he misses it, it was probably a random error. If he gets it right, it tells us he's not low ability, which we already knew.

To understand the importance of matching the difficulty of the item to the ability of the test taker, let's see what happens when we give two versions of a test to a group of test takers whose abilities range from average to well above average. We will call these two versions of our test Version 1 and Version 2. Version 1 has items ranging in difficulty from fairly easy to extremely difficult. The medium ability test takers would miss most of the difficult items (due to lack of ability) and have the lowest scores of the group. The high ability test takers would answer almost all of the items correctly and have the high scores. Random errors would occur, but they play a minor role in why some people have low scores and others have high scores. This is how things should go.

Version 2 of our test contains only easy items. It should be obvious that all of the test takers have enough ability to answer every one of these items correctly – and they will, except for items they miss due to random errors. As a result, just about every test taker will have a high total score on the test. Worst of all, most of the differences among the scores will be due to random error and not due to ability differences among the test takers. The highest scoring test takers have the highest scores not because they had the highest abilities but because they were lucky in the random error department. I hope this sounds bad, because it is. Think back to Chapter 1. One of the goals of testing is to give people with different ability levels different scores. Version 2 of the test has failed to do that. All because of a poor match between item difficulty and test taker ability.

You might be wondering, how did we know the ability of these test takers before they took the test? (And if we already knew their ability level, why are we bothering giving them the test?) Well, most of the time we don't know the ability of the test takers. But there are some occasions when we can make an imprecise approximation of their ability level. If we have a group of first grader students, odds are good that they are of low ability on just about everything compared to general society. Similarly, a group of advanced college students would be of high ability for just about everything in the academic realm as compared to general society. What if we don't know squat about the ability of our test takers? The best items would be of mostly average difficulty with some hard and some easy questions added to the mix. That is, we'll assume the ability of the test takers is normally distributed, mostly average, with a small amount of high and low ability people as well. The right items for these people will be similarly distributed in terms of difficulty.

Now let's talk statistics. We only have one to deal with and that's the difficulty of the item,

called *p* value. Now this *p* value isn't the same as the *p* value from statistics class which related to testing the null hypothesis with significance tests like *t* tests and ANOVAs. This *p* value simply tells us the percent of people who answered the item correctly (i.e., in the keyed direction). Very simple to compute. (Interesting note: I didn't use the term *p value* in Chapter 8, but the concept, in the form of *percent correct*, was used repeatedly. Go back and see if you can find it.) Obviously, *p* values are limited to dichotomously scored items (i.e., scored as right or wrong, or stated more generally, answered in the keyed direction or not). If our items have more than two possible scores, like the common multi-point Likert item (e.g., "I like cats." "Strongly Disagree, Disagree", etc.), we won't be able to compute *p* values. In that case we can compute the item mean and use it for the same purpose. High means indicate attitudes with which most agree and low means indicate attitudes with which most disagree. It's just not as cool. Getting back to dichotomously scored items, let's compute some *p* values for the data below.

Person	Item 1	Item 2	Item 3
Carla	1	1	0
Michael	1	0	0
Jane	1	1	1
Roger	0	0	0
Bertrand	1	1	0

In order to compute *p* values the data must already be scored. And this dataset has been scored. How do I know? I made it up, and I'm saying it has been scored. Now to compute the *p* values for first item, we note that 4 out of 5 people answered correctly. Thus the *p* value is 4/5 or .8. It's an easy item. Most people are answering it correctly. For Item 2, 3 out of 5 answered correctly, so $p = .6$. Close to average difficulty. Finally, Item 3 is difficult. Only one person answered correctly, $p = .2$.

How do we use these *p* values for item analysis? If you wrote a test a large number of items with high *p* values (let's say, 1.0), then these items are not separating your test takers. They all get them correct (same problem if the *p* values are all 0.0 – everyone missed the items). It's like adding a constant to everyone's score. That doesn't help you assess their standings on the construct. Moreover, you can't correlate a constant with anything. So performance on these items is irrelevant to other variables. Here's another thing: Items with 1.0 or 0.0 *p* values have zero variance (*zero-variance items* is sort of a nickname for these items). Let's put all this together. Remember composite variables from the previous chapter? The two determinants of the variance of a composite variable are (a) the variance of the items that comprise the composite and (b) the item intercorrelation. Items with *p* values of 1.0 or 0.0 have zero variance and don't correlate with the other items. Which is a long way of saying what we already

REVIEW 1 Difficulty-Based Item Analysis

Question 1 of 4

What is the *p* value of an item that is answered incorrectly by 82 out of 100 test takers?

- **A.** .82
- **B.** .18
- **C.** Something else

Check Answer

said: Items with 1.0 *p* values or 0.0 *p* values do very little for us as far as determining a person's standing on the construct.

OK, you say. I've got it. We don't like items with *p* values of 1.0 or 0.0. But those must be rare. What about items with .93 or .02 *p* values? The answer is that the problem is the same, just to a reduced degree. Those items do very little for us in terms of measuring a person's standing on a construct.

There you go, that's difficulty-based item analysis in all its glory. Now the best application of this will ultimately be computer adaptive testing, but we'll save that talk for another day. There is one other issue related to this that we will discuss a bit later in this chapter. That's coming when we get to item analysis for domain-referenced tests.

Internal Consistency and Empirical Keying Item Analysis Overview

The next two types of item analysis have something in common. For both empirical keying and internal consistency item analysis a good item is one that correlates well with some variable. The only issue up for grabs is: What is this variable with which we are correlating the items? That's where the two methods differ. Empirical keying correlates items with a variable independent of the test, whereas internal consistency correlates items with the other items on the test.

Internal Consistency Item Analysis

With an internal consistency item analysis, a good item is one that correlates well with the other items on the test. This method of item analysis has everything in common with internal consistency estimates of reliability. Recall from Chapter 5 that the basic split-half procedure involved divid-

ing the items on a test into two groups, getting a total score for each half of the test, and correlating the scores. Things were good if the correlation was strong, meaning that (a) there was little random error and (b) the factors that caused people to get high scores on one half of the test were the same factors that caused people to get high scores on the other half. That is, each half of the test measured the same thing. Ultimately, we used coefficient alpha to compute split-half reliability estimates because it solved most of our problems. Getting back to internal consistency item analysis, we'll keep items that correlate well with the other items. Doing so will have the ultimate effect of giving us a high coefficient alpha (and alpha is sort of like an average correlation among the items).

Our statistic of choice will be the corrected item-total correlation. Forget the *corrected* part for a second. An item-total correlation is the correlation between the scores on a given item (let's say Item 1) and the total score on the test. A strong, positive item-total correlation indicates that people who do well on Item 1 also do well on the other items (and vice versa). No guarantees here, but this also suggests that Item 1 may be measuring the same construct as the other items. We call this statistic a *corrected* item-total correlation because when we correlate Item 1 with the total score on the test, we don't want Item 1 to be a part of the total score. If Item 1 was a part of the total score, it would be a part of both sides of the correlation, inflating the correlation. Thus, on a 20-item test, a corrected item-total correlation for Item 1 is the correlation of Item 1 with the total score on Items 2 through 20. If Item 1 has a low corrected item-total correlation (say, equal to 0.0), then we don't want it on our test. If its item-total correlation is strongly negative (r = -.4 or so), we really don't want it on our test unless we can determine that a scoring error led to the negative sign. In such a case, we would fix the scoring error and rerun the analysis. But that's just Item 1.

What about the rest of the items? We need to compute corrected item-total correlations for Item 2-20. The corrected item-total correlation for Item 2 would be the correlation between scores on Item 2 and the sum of Items 1 and 3-20. In short, you could say that a corrected item-total correlation is the correlation between Item *k* and the total score of the rest of the items (i.e., all but Item *k*).

Let's talk about an item that has a 0.0 corrected item-total correlation. As I said, we don't we want it on our test. But why? There are two possible reasons why our item could have such a low correlation with the other items. The first is that the item is full of random error. As you recall, randomness doesn't correlate. So that would explain the low correlation. The other reason is that maybe the item is mostly free from random error, but it measures a different construct from the rest of the test. Given that this item analysis technique is about internal consistency, we don't want to keep items that measure other constructs. (I know I said that there were two reasons, but there is actually one more possibility. Maybe our item doesn't have a random error problem, but the rest of the items do. Consider a 20-item test with 19 unreliable items and one reliable item. The one reliable item wouldn't correlate with the total score on the other 19.)

Now when we do our analysis, we start with all of the items and delete them one at a time. Why not two at a time? Because every time we delete just one item, the list of "other items on the test" has just changed. For example, the corrected item-total correlation for Item 4 is the correlation between Item 4 and the sum of Items 1-3 and 5-20. But if we delete Item 2 because it's a loser, the corrected item-total correlation for Item 4 is now the correlation between Item 4 and the sum of Items 1, 3, and 5-20.

As mentioned earlier, the effect of deleting items that fail to correlate well with the other

items is that coefficient alpha for the test will be maximized. So we need to check alpha during this analysis. We should do this at every step because ultimately, there will be a point where the deletion of additional items no longer increases alpha by a substantial amount. It's that second to last word *substantial* that's important. We can almost always get alpha a little higher. A trivial bit higher. But we have to ask ourselves, is it worth deleting another item just to get alpha .0001 higher? The answer is no. What if we can get alpha to go from .60 to .65 by deleting Item 3? In that case, I would delete the item. I recommend a .04 or .05 threshold. If coefficient alpha will increase by at least .04 or .05 (depending on how strict you want to be), then throw out the item. Anything less, and we'll keep the item.

Enough talk, let's see some data in action. What follows are the results from an analysis of an eight-item test. The items have been scored and the initial coefficient alpha is .24. The corrected item-total correlations and a column titled "Alpha if Item Deleted" (which, as the name suggests, tells us what coefficient alpha will be if we delete the item) are below. We can use this to determine if we should bother deleting the item. Remember that we are starting with an alpha of .24.

Item	Corrected Item-Total Correlation	Alpha if Item Deleted
1	0.05	0.24
2	0.15	0.18
3	0.23	0.13
4	-0.23	0.37
5	0.27	0.09
6	-0.05	0.32
7	0.10	0.22
8	0.25	0.12

Clearly, there are big problems with Items 4 and 6. They both have negative corrected item-total correlations. I've checked, and there aren't

any scoring errors. Thus, these are the first items we will target for deletion. Remembering that we do this one item at a time, we pick Item 4 first because it is more strongly negative than Item 6. Also note that if we throw it out, alpha will increase to .37. That's quite a bump up.

So, Item 4 is gone, here are the new results. Our coefficient alpha has increased to .37, just like they promised.

Item	Corrected Item-Total Correlation	Alpha if Item Deleted
1	0.05	0.40
2	0.13	0.35
3	0.23	0.30
4	–	–
5	0.34	0.22
6	-0.04	0.46
7	0.18	0.33
8	0.31	0.25

Now Item 6 has the worst corrected item-total correlation. Deleting it will raise alpha to .46. Out it goes.

Item	Corrected Item-Total Correlation	Alpha if Item Deleted
1	0.10	0.49
2	0.21	0.43
3	0.26	0.40
4	–	–
5	0.40	0.31
6	–	–
7	0.16	0.45
8	0.25	0.40

Alpha did indeed increase to .46. Which item has the worst correlation with the other items? Item 1 you say. Should we delete it? No, you say. Why, I ask? Because, you say, if we delete Item 1, our coefficient alpha will increase by only .03 units (from .46 to .49). Good call, I say. Then our inter-

nal consistency item analysis is done. We have an eight-item test that is as internally consistent as we could get it to be. Although a coefficient of internal consistency of .46 is fairly lousy, it is the best we could do for these items. That's the way it goes sometimes.

Empirical Keying Item Analysis

In an empirical keying analysis, a good item is one that correlates well with an external variable. What is an external variable? It is any variable that is not a part of the item pool. Let's say that we wrote 43 items measuring math ability. An external variable is any variable other than these 43 items. And it could even be on the same piece of paper. We could ask people when they finish the test to write their ACT Math Composite score at the bottom of the test. Our item analysis would simply be the correlation of each item with the ACT scores. We would keep the items that correlate with ACT scores. At the end of this process, we would have a test that is associated with ACT scores and could be used to predict ACT scores if we so desire.

Really, that's all there is. As before, it would be wise to score our items before we start our analysis. Also, if we encounter a strong negative correlation (when the rest of the correlations are positive), we should check to see if our scoring was correct. Of course, if we have one positive correlation among a field of negative ones, we need to check the accuracy of the scoring of the one positive correlation. Here we go. Let's say that we have a five-item test consisting of knowledge about some movie. It could be any movie. Just imagine that it's your favorite movie. (I'm imagining that it's *Citizen Kane*. It's not really my favorite movie – I just tell everyone that it is.) The external criterion is a measure of how often they have seen the movie. Clearly, people who have seen the movie multiple times should know a great deal of information about the movie.

Item	Correlation with External Criterion Variable
1	-0.29
2	0.55
3	0.55
4	0.37
5	0.07

We note that Item 1 has a fairly strong negative correlation. We check for scoring errors and can't find any. Thus, Item 1 is out. (Unlike internal consistency item analysis, we can delete multiple items at once with an empirical keying item analysis.) We also note that Item 5 has a very weak correlation and decide to throw it out. These two items were no-brainers. What standard do we use to determine whether a correlation is strong enough? There are a number of answers to that question that depend on a number of issues. We'll use a fairly liberal .2 standard. That is, any item with a correlation weaker than .20 will be deleted. In our case, we are done deleting items as Items 2-4 all are greater than .20. So, the final version of our test consists of three items (Items 2-4), all of which predict the external criterion variable fairly well. How well does our three-item test as a whole predict the criterion variable? By that, I mean, how well does the total score on our three-item test predict? To answer that, I formed a total score consisting of the total of scores on the good items (Items 2-4). The correlation of the total score with the external criterion variable is .62. That's some pretty good prediction. Now, we'll still need to do a validity study with a new sample of data to verify this, but we're obviously off to a good start. You know what's interesting? The single best item (Item 3) had a correlation of .554. But when we added Items 2 and 4 to it to form the total score, the correlation of the total was .62. That is, a collection of good items works better than any one item by itself. Yet another reason to not have one-item tests.

Item Analysis for Domain-Referenced Tests: Content Validity

Overview. Just as a reminder, the point of a domain-referenced test is to determine if a person's score meets a set standard. The score is not compared to other people's scores, but rather to a set performance level, as quantified by a cutoff score. If the test taker's score is greater than the cutoff score, then the test taker passes the test. If not, they don't.

Although it's possible to draw a domain-referenced interpretation of a score from any type of test, it's a bad idea for tests other than content valid tests. What's a content valid test? A test is content valid if the behaviors measured on the test are a representative sample of the behaviors in the content domain. And the content domain is the set of all behaviors relevant to the purpose of the test. So the test is the small thing, and the content domain is the big thing. The test taker's performance on the test is generalized from the small thing (the sample of behaviors on the test) to the big thing (the content domain). As an example, if you can perform the twenty or so behaviors on the road part of the driver's test, then we infer that you can perform all of the behaviors involved in driving a car. Does this actually work? It does if the test content is a representative sample of the content domain and if the content domain has been defined correctly. Thus, the goal of our item analysis procedure is to develop a test that is content valid.

Content validity will be discussed in detail in Chapter 10. For the present purposes of item analysis, we'll just discuss the basics. Furthermore, let's limit the application of the content validity strategy to tests designed to determine a person's performance on a narrowly defined set of observable behaviors (like driving a car). Application of the content validity strategy to tests of unobservable constructs like depression will also be dis-

cussed in Chapter 10 (Spoiler alert: The content validity strategy doesn't work well for tests of unobservable constructs).

Procedure. A content validity study proceeds as follows.

I. Define the content domain.
II. Make the test.
III. Establish that the behaviors and/or information measured on the test are a representative sample of the content domain.

How is this done? The content domain is defined by a careful analysis of the target of the test. If it's a driving test, then we study driving. We document not only the behaviors performed but also the information required to drive. That's the first step. For the second step, consult Chapter 6. The final step (and this is the item analysis part) uses an expert judgment process. Experts render their opinion regarding the representativeness of the test content in the final step.

How do the experts decide the issue? They simply compare the content of the test with what's in the content domain, then ask three questions. Is there something in the content domain that is not on the test? (If so, then write an item to measure that behavior.) Is there something on the test that is not a part of the content domain? (If so, then drop the item.) Are the items on the test a representative sample of the content domain? (If not, then write additional items to address the underrepresented content areas. Or delete items from overrepresented content areas.) There's your item analysis.

In summary, our goal is to develop a test that tells us how well a person can perform a given activity (or how much they know about a narrowly defined topic). We list all the behaviors and/or information associated with that activity (the content domain). We write a test that contains a representative sample of these behaviors. This test is intended to be a miniature version of activity we

want to measure. As long as the content domain has been accurately and comprehensively defined and if the test is a representative sample of this domain, then we can validly generalize from test performance to performance on the larger domain.

Final Thoughts on Content Validity. If you're like me when I learned about content validity, there's still something bothering you. Where are the correlations? There aren't any. What statistics do we have for content validity? Mostly, we have agreement statistics, which quantify how well our experts agree whether a given test item measures a behavior in the content domain.

Final Thoughts on Item Analysis

So that's four kinds of item analysis, which if independently performed on an item pool, would yield four different collections of items. Their goals are different; thus, the items that they retain are different. Chart 2is a handy summary of the various definitions of a good item for each of the item analysis techniques.

CHART 2 The Nature of a Good Test Item According to Various Item Analysis Techniques

Item Analysis Technique	A good item is...
Difficulty-Based	...one with a difficulty that matches the ability of the test taker.
Internal Consistency	...one that correlates with the other items.
Empirical Keying	...one that correlates with a variable independent of the test.
Content Validity	A good set of items are those that are a representative sample of the content domain.

Validity

10

Interpretations and more interpretations.

The Definition of Validity

We said it in Chapter 4, and we'll say it again: A reliable test *may* be a good test. An unreliable test (which measures nothing but random error) is clearly bad. OK, so a reliable test does not measure random error, but does it actually do what we want it to do? That is the validity question. What is validity? Well here's the definition offered by *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 9):

> Validity is "...the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test."

That's a long way of saying that validity is all about having evidence to support the interpretations we draw from test scores. It is not about the test itself, or the scores. It's about how we interpret the scores (many people state this as "inferences we draw from the scores," that's fine). What are some interpretations? Let's say that we take a test that is supposed to be an IQ test. And we get a high score. What is the interpretation of the score? We interpret it to mean that we are smart. Good for us. Now let's say we take the ACT, a test designed to predict college performance, and we get a high score. In this case, the interpretation is that we will succeed in college. Now let's say that we take an integrity test. An integrity test has questions that ask you if you have stolen from previous employers. Now let's say we get a low score (and low is bad). The interpretation is that we will steal from our potential employer. Or we could take a driver's licensing test and get a high score. The interpretation is that we will be able to drive a car at some minimum level of competence.

There are many possible ways to interpret a test score. We could interpret a high ACT score to mean that a person is smart. We could interpret a low score from an integrity test to mean a person is depressed. We could interpret a high score on

an IQ test to mean that a person will be a good driver. But do we have any evidence to support these interpretations? If not, then those interpretations are not valid.

So how do we gather evidence to support a given interpretation of a test score? We do a validity study. There are three types of validity studies, a criterion-related validity study, a content validity study, and a construct validity study. Note that these are not three types of validity, but rather three different ways we can gather evidence to support the validity of the interpretations of scores that we draw.

Criterion-Related Validity

Criterion-related validity is all about predictive inferences. Do scores on this test predict some criterion variable? This is how the ACT is used. ACT scores are used to predict college performance. Thus, we interpret a high ACT score to mean that a person will succeed in college. The basic plot of a criterion-related validity study is fairly simple. First, determine what it is you want to predict (college performance, job performance, driving performance, relapse rate, etc.). Second, give the test to a group of people. Third, measure performance on some criterion variable for these same people. Finally, correlate the scores on the test with scores on the criterion variable. If we have a strong and significant correlation, we have evidence that our test predicts performance on this variable and can interpret a high score to mean that a person will succeed.

There are two experimental designs for a criterion-related study: the predictive design and the concurrent design.

Predictive Design. With a predictive study, we give the test to a group of people who have not had any experience at the thing we want to predict. For example, if we are trying to predict col-

lege performance, our sample will be composed of people who have not yet attended college. High school seniors make for a fine group in this case. After they take the test, we give them some experience at the thing we are trying to predict. Sticking with the college example, we let them attend college. We probably won't be able to let everyone into college, so at this stage, we are going to have to make some selection decisions. Once the selected group is in college, we need to wait while they experience it. Some will succeed. Others will fail. We give them a chance to do that. After we have waited a sufficient amount of time, we measure performance on our criterion variable. Once we have done so, we can correlate scores on the test with scores on the criterion variable. To summarize the predictive design:

1. Determine what we want to predict.
2. Give the test to inexperienced people (i.e., applicants).
3. Make selection decisions.
4. Wait while the selected group gains experience on the behaviors relevant to the criterion variable.
5. Measure criterion performance.
6. Correlate the scores.

The hidden danger is in Step 3. How we make our selection decisions can adversely impact the correlation that we obtain in Step 6 (called the validity coefficient). The optimal scenario would be to select at random. For practical reasons, that's not likely to happen. The thing we want to avoid at all costs is making the selection decisions on the basis of the scores of the test that is being validated. Doing so will introduce direct range restriction, which will weaken the validity coefficient we obtain. A discussion of how this happens is in order.

Interactive 1 presents an explanation of how range restriction weakens a correlation.

What are we to do about range restriction? Obviously we should avoid it to the extent possible. But how? Ideally, we would use a random process to make the selection decisions. If that's not possible, we can use an alternative test (i.e., a second test) to make the selection decisions while we still collect data on the test we are validating. With this second test, we will still have range restriction (called indirect range restriction), but it's impact on the correlation is likely to be minimal.

INTERACTIVE 1 How Range Restriction Weakens a Correlation

Concurrent Design. One problem with the predictive design is the need to wait. If we are predicting college performance, we will have to wait at least one semester, and probably more. Ideally, we want our measure of performance to be based on a large sample of behavior, and one semester of college (in our ACT example) is not much. We probably want at least two years of performance before we measure college performance. Thus, we might have to wait two years. That's a long time. Can't we get the results any faster? Yes, if we do a concurrent study. In a concurrent study, we will give the test to experienced people. That is, people who already have experience at the thing we are trying to predict. Because they already have experience, we don't need to wait, we can go right to the part where we measure criterion performance. So the steps are:

1. Determine what we want to predict.
2. Give test to a group of experienced people (i.e., current employees).
3. Measure criterion performance.
4. Correlate the scores.

Much quicker this way. And we don't have to make making selection decisions. They've already been made. But there's nothing for nothing. New problem: The people taking the test may not be all that motivated. In a predictive study, the people taking the test usually want something: a job, college admission, etc. And they think that the test is their ticket to get in. So, they are always trying hard. But in the concurrent design, we've got people who already have experience. Their futures are not hanging on the outcome of the test results. In short, there's nothing in it for them. This lack of motivation can affect results, which ultimately distorts the validity coefficient. But, you say, couldn't we offer them rewards and incentives to get them to try harder? Yes, and that may solve some of the problem.

Problems with all Criterion-Related Validity Studies. We've got more than a few other problems to deal with. First, because all criterion-related validity studies end up with a correlation, anything that is a problem for a correlation is a problem for our study. That includes range restriction (mentioned above when we talked about how the selection decisions were made), sampling error, and unreliability of measurement. Small sample sizes are greatly affected by sampling error, resulting in deviant correlations, either too high or too low. And it's not like we know this when we compute them. The correlation doesn't come with a label stating, "This correlation is .23 units greater than the population value." Nobody loses sleep over correlations that are too high, but the ones that are too low have a bad habit of being non-significant. Solution: Use large sample sizes. Finally, variables that are not perfectly reliable will lower our validity co-

efficient. Solution: Use highly reliable measures of our predictor and criterion variables.

The Criterion Problem. There's one more item on my list of problems. It is called the Criterion Problem, capital letters and all. The problem with The Criterion Problem™ is that we will never obtain a perfect measure of the criterion variable. As such, the correlation between scores on our test and our imperfect measure of the criterion variable doesn't indicate how well the test actually predicts criterion performance. As the measurement of the criterion variable worsens, our validity coefficient becomes more misleading. It is possible that we can obtain a strong and significant validity coefficient (e.g., $r = .5$) for a test that is not really valid. Why? Although our test predicts this flawed measure of the criterion variable, it does not predict the way people actually perform on the job (or in school, etc.). All we got was a bad measure of criterion performance. Conversely, we could obtain a 0.0 correlation with a terribly flawed measure of the criterion variable, but our test may actually be valid. How? What if our measure of the criterion variable is all random error? Well, nothing will correlate with it. Thus, $r = 0.0$ for every test that we use to predict criterion performance. What's the moral of this story? Make sure that your measurement of the criterion variable is as good as possible. Don't just grab a convenient measure of criterion performance. Get a good one. Or don't do the study.

Final Thoughts. It should be clear that there are a lot of ways for a criterion-related validity study to go wrong. A small sample could introduce too much sampling error. One or both variables could be unreliable. We could measure the criterion variable poorly in other ways. There could be range restriction. In a concurrent design, the test takers may not be motivated in the same way real test takers would be. It all seems like too much to bear. Why would anyone bother trying to gather validity evidence with an approach so prone to

problems? The first answer is, if you're trying to validate a predictive inference, criterion-related validity is pretty much the only game in town. The second is that when done correctly, a criterion-related validity study provides persuasive evidence that is difficult to discount. Obviously, the trick is in doing the study correctly.

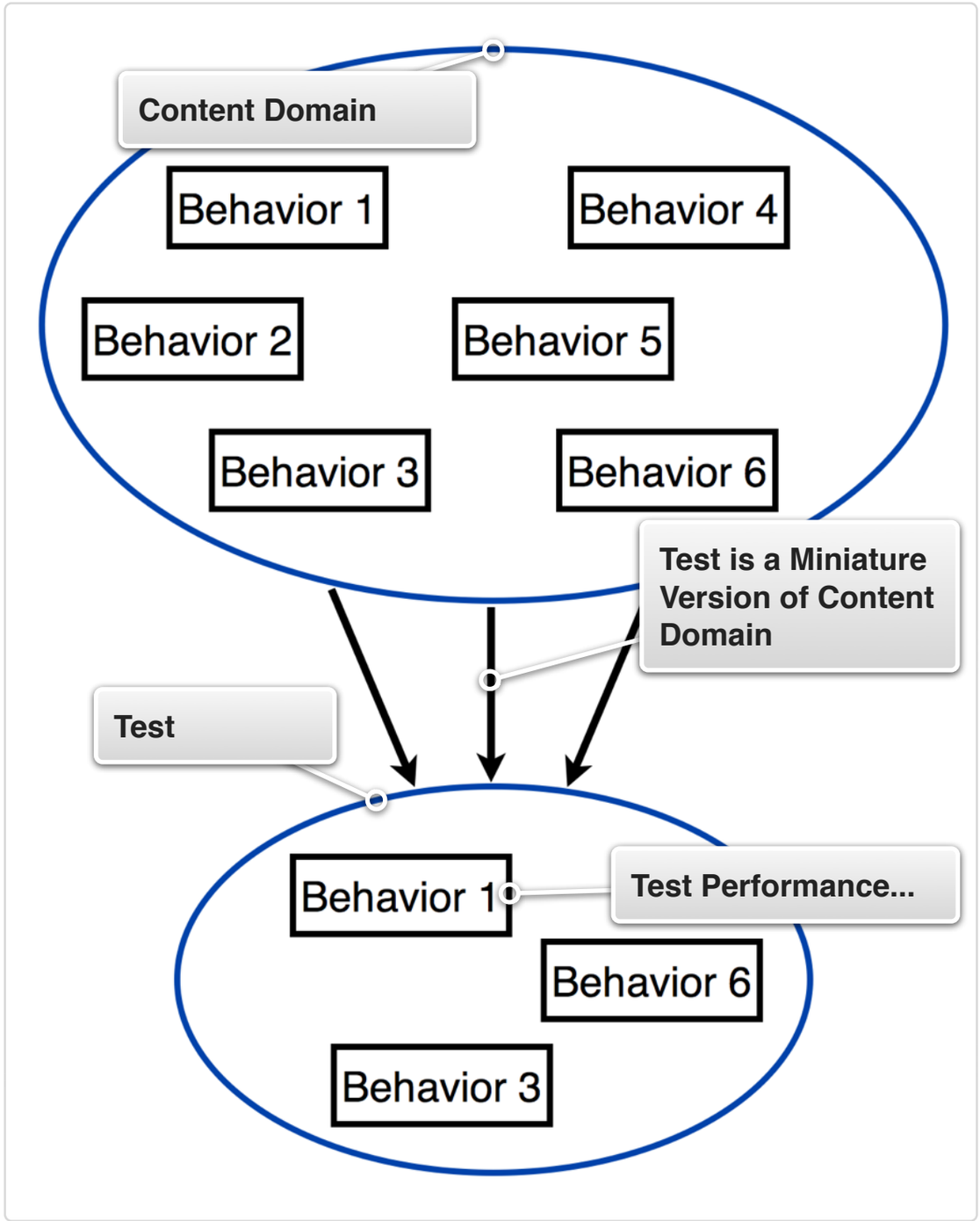
Content Validity

Definition. A test is content valid if the behaviors measured on the test are a representative sample of the behaviors in the content domain. The content domain is the set of all behaviors relevant to the purpose of the test. The controversial issue with content validity relates to the various purposes we can have for a test. For now, let's limit the purpose to performance of an activity (e.g., driving a car) or knowledge of a narrowly defined topic (e.g., state capitals).

We've talked about inferences in previous discussions of validity. What's the inference with content validity? The inference is a generalization drawn from the test taker's performance on a small thing (the sample of behaviors on the test) to a big thing (the content domain). As an example, if you can perform the ten or so behaviors on the road part of the driver's test, then we infer that you can perform all of the behaviors involved in driving a car.

Does this actually work? It does if the test content is a representative sample of the content domain and if the content domain has been defined correctly. There are also some other conditions that need to be met. We'll get into those later. First, let's talk about how we conduct a content validity study.

FIGURE 1 The Concept of Content Validity



Procedure. A content validity study proceeds as follows.

I. Define the content domain.
II. Make the test.
III. Establish that the behaviors and/or information measured on the test are a representative sample of the content domain.

How is this done? The content domain is defined by a careful analysis of the targeted activity. If it's a driving test, then we study driving. We document not only the behaviors performed but also the information required to drive. That's the first step. For the second step, consult Chapter 6. The final step uses an expert judgment process. Experts render their opinion regarding the representativeness of the test content in the final step. It is in this third step of the process where one can say that the test has been validated. How do the experts decide the issue? They simply compare the test content with the content domain and address

three questions. Is there something in the content domain that is not on the test? Is there something on the test that is not a part of the content domain? Are the items on the test a representative sample of the content domain? Once these issues are addressed, you have a content valid test, assuming some other requirements are met.

What are these other requirements? Well, as Pedhazur and Schmelkin (1991) wrote, "Validity refers to inferences made about scores, not to an assessment of the content of the instrument" (p. 79). What they mean is that the process given above for establishing the content validity of a test is really just a process for creating a test. And, as we said earlier, validity is about interpretations of test scores, which extends far beyond the content of the test. Missing from the process described above are standardized procedures for administering the test, scoring the responses, and interpreting the test scores. The good news is that the first and the last components are rather easy (there's only one permissible interpretation to draw with a content valid test, the generalization interpretation), and the second one should also be easy for almost any performance domain assessed with a content valid test. In summary, given standardized and logical procedures for administering, scoring, and interpreting scores on a test constructed with a content validity strategy, one has adequate support for the interpretation that performance on the sample of behaviors measured on the test generalizes to performance on the larger domain of behaviors.

I hope that everything is clear up to this point. Our goal is to develop a test that tells us how well a person can perform a given activity (or how much they know about a narrowly defined topic). We list all of the behaviors and/or information associated with that activity (the content domain). We write a test that contains a representative sample of these behaviors. This test is intended to be a miniature version of activity we want to meas-

ure. As long as the content domain has been accurately and comprehensively defined, and as long as the test is a representative sample of this domain, then we can validly generalize from test performance to performance on the larger domain (the activity). In this case, the content validity process works, and the reason it works is that every step of the process deals with observable behaviors (and/or specific information that is traced back to the activity).

Controversy. Let's say I want to develop a test of an unobservable construct like depression. Can I use the content validity strategy to validate such a test? It might help to review the two definitions of construct from Chapter 1. Everything we've discussed up to this point has been about the second definition of construct, the *set of behaviors* definition. We are now switching gears and addressing the first definition of construct, the *unobservable cause* definition.

Applying the content validity process to an unobservable construct complicates matters because things are fundamentally different. We are now trying to evaluate whether the observable behaviors measured on the test are a representative sample of something that is unobservable. If you want this problem in the form of a question, here it is: if something is unobservable, how do we know if a set of behaviors (the test) is a representative sample of it? Do our experts have special powers that mere mortals lack?

Hardcore fans (and there are some) of applying the content validity process to tests of unobservable causes have a response to this question. Their response is, "The experts are simply comparing one observable thing (behaviors on the test) with another observable thing (the behaviors in the content domain)." True, but this assumes that the construct was correctly defined as a set of observable behaviors (i.e., the content domain). And that's the fundamental point of concern with this

approach. The construct is unobservable, so there's no way to know if it has been correctly and comprehensively defined as a set of observable behaviors. All arguments regarding this issue are just arguments about definitions of something unobservable. Hence, every position is unprovable. (This is no minor issue. Arguments rage not only regarding the definition of a construct, but whether the construct even exists.) Can we comprehensively define something unobservable in terms of observable components? (Moreover, how would we know if we did it correctly?) If the answer is no, then the content validity process is not appropriate for these sorts of constructs. The answer is no. It's not all bad news for tests of unobservable constructs; the construct validity strategy is perfect for these tests.

Content Validity Applied to Three Tests. Let's see how well the content validity process works when applied to three different types of tests: the road portion of a driving test, a depression test, and a scholastic achievement test. First, the road portion of a driving test. Here, the construct is a set of driving behaviors. Driving is a set of observable behaviors. Just list them. That's your content domain. It's work, but it can be done. Now make a test that is a sample of behaviors from this domain and check to see that it is a representative sample of the domain. Open and shut. All very simple. Observable behaviors (or specific information) at every stage. We can generalize from performance on the test to performance on the entire domain of car driving behaviors.

Next, a depression inventory. Depression is an unobservable construct, so this will be a problem. The first step is to define the content domain where we list all of the behaviors related to depression. This needs to be a fully comprehensive list. As we discussed, we can never know if we have accurately and comprehensively defined something unobservable. So, the content validity process breaks down here. We can't generalize test per-

formance because we can never know if our test is a representative sample of the larger domain. Thus, the content validity strategy doesn't work for this type of construct. One possible solution: What if you used the DSM definition of depression as your content domain? The DSM lists a number of behaviors associated with depression. You could write a test that is a representative sample from this list. Everything obtained from the DSM onward is observable. Sounds good, but it assumes that the DSM's list is accurate and comprehensive. So you couldn't say that you have a content valid test of depression, but you could say that you have a content valid test of depression *as defined by the DSM*. That's something. But you're still better off with a construct validity study.

And finally, the third test is a traditional, grade school achievement test. Can the content validity strategy be validly applied to this test? It might appear to be that this is another case of a test of an unobservable construct. Is it? We'll answer that with a question: what is the content domain for an achievement test? Answer: everything that was taught during the school year. Is this an unobservable construct or a set of observable behaviors? It's similar to a set of specific information (which we treat like a set of behaviors). With an achievement test, we want to know if the student has learned Topic X (state capitals), Topic Y (names of rivers), and Topic Z (names of mountains); we do not care about some underlying ability. The content domain can be listed in same way that all of the driving behaviors can be listed (the content domain is every topic covered in class). Thus, as with the road driving test, everything is observable. In fact, we see this very thing with the written driving test. The written portion of the driving test is supposed to be a representative sample of everything a driver has to know to drive a car. We can comprehensively define the content domain. Thus, we can determine whether the test is a representative sample from this domain. So, it's

all good. (If you think about it, the generalization inference of content validity is really behind every test given in a college class. Whether these tests are actually content valid is another story. That depends on whether the instructor took the time to insure that the test is a representative sample of the course material.)

To summarize, the content validity strategy is well suited for inferences regarding generalizations from a sample of behaviors to the larger domain of behaviors. It is not suited for generalizations to an unobservable construct.

Face Validity. One last content validity issue. Content validity is not face validity. A face valid test is a test that appears to be valid to the test taker. What kinds of tests do test takers think are valid? The ones where the format of the test matches the format of whatever the test is used for. So let's remember face validity like that. A face valid test is a test in which the format of the test matches the format of the performance domain. For example, a face valid test for the job of forklift operator would involve driving an actual forklift. A paper and pencil test for forklift operator in which a person merely answers multiple choice questions about forklift operating principles would not be face valid. Face validity and content validity are two separate issues. It is possible to have a highly face valid test that is not content valid (just drive the forklift in a straight line; that's it). It is also possible to have a content valid test that is not face valid (once again, the paper-and-pencil test of forklift operating principles).

A common reaction to learning about the difference between face validity and content validity is to say that face validity doesn't matter. That is, the format of the test doesn't matter. Is that true? Is the format of the test completely irrelevant to the content validity of the test? Consider the following scenario. We want to make a content valid test for that forklift operator job we mentioned. We

analyze the job and make a list of all the things a forklift operator has to know in order to perform the job. We make a multiple choice, paper-and-pencil test of these concepts, and this test is a representative sample of the job. Problem: The job doesn't involve reading, and you obviously must read to pass our paper-and-pencil test of job knowledge. What we have done is measured something (reading ability) that is not a part of the performance domain. Now, you might say that anyone who can't read really shouldn't have this job. But why? Reading isn't a part of the job. And even if it is, what if the reading level required for the job is low (third grade), but the reading level of the test is high (tenth grade)? We are still measuring an irrelevant behavior. Not good. So we must always be cautious with *how* we measure the behaviors/information on our content valid test. It should also be mentioned that even if our paper-and-pencil test of forklift operating principles didn't have reading-level problems, our test still isn't content valid for the simple reason that a written test is a knowledge test and successful performance of the job requires more than mere knowledge about how to operate a forklift. It requires actual operation of the forklift. A person may know a great deal about how to perform a behavior but be unable to successfully perform it.

Final Thoughts on Content Validity. If you're like me when I learned about content validity, there's still something bothering you. Where are the correlations? There aren't any. Correlations are a part of criterion and construct validity but not content. What statistics do we have for content validity? Statistics related to agreement among the experts. Also, reliability is pretty important for the same reasons that reliability is always important. But wait, if we don't correlate the test scores with criterion scores, how do we know if the test is valid? The answer is that we are not making a prediction inference. We are making a generalization inference. If the test content is a representative sample

of a larger domain of behaviors, then performance on the test can be generalized to performance on the entire domain of behaviors.

Construct Validity

With construct validity, we are concerned with whether the test measures the construct it is supposed to measure. That is, if my test is supposed to measure intelligence, can I correctly interpret a high score as indicating that a person is smart? Anyone can say that their test measures intelligence (or schizophrenia, or depression, or what have you), but how do we know if it does? In order to support our claims of the construct validity of our test, we need to demonstrate two things: convergent validity and discriminant validity.

Convergent validity is when scores on our test correlate strongly with scores on another test measuring the same construct. That is, people who score one way (e.g., high) on our test also score the same way (high) on another test of the same construct (in short, we get a strong, positive correlation between the two). The logic is simple. If our test really measures intelligence, then scores on our test should correlate strongly with scores on an existing test of intelligence. The steps to computing convergent validity coefficients are short and sweet.

1. Give our test to a group of people.
2. Give another test measuring the same construct to the same group of people.
3. Correlate the scores and hope for the best (a strong correlation).

Discriminant validity is the opposite of convergent validity. In fact, discriminant validity is unique in that it is one of rare times in which we hope to obtain a zero correlation. Nothing would thrill us more than a 0.0 correlation. The logic is as follows. If our test measures construct X, and construct X is unrelated to construct Y (in a con-

ceptual sense), then scores on a measure of construct X should not correlate with scores on a test of construct Y. That is, scores on our test should not correlate with scores on a test measuring a different construct. The steps are almost the same as before.

1. Give our test to a group of people.
2. Give another test measuring a different construct to the same group of people.
3. Correlate the scores and hope for the worst (a weak correlation).

If the test displays adequate discriminant validity and convergent validity, then we have pretty good evidence that it measures the construct in question and not some other construct (assuming something we'll discuss in a page or two). We can validly interpret a high score as indicating a high standing on the construct in question.

The data from a construct validity analysis can be displayed in a matrix of correlations. This matrix is called a Multi-Trait Multi-Method matrix (or MTMM matrix). Elements in a MTMM matrix include convergent validity coefficients, discriminant validity coefficients, and, optionally, reliability coefficients. The word *trait* is a synonym for construct and the word *method* can be thought of as a type of test. However, I find it more useful to think in terms of *methods of measurement* instead of test since a single test can measure multiple constructs. Below is a MTMM matrix displaying data from two constructs (depression and schizophrenia) measured two different ways (paper-and-pencil test and observational checklist).

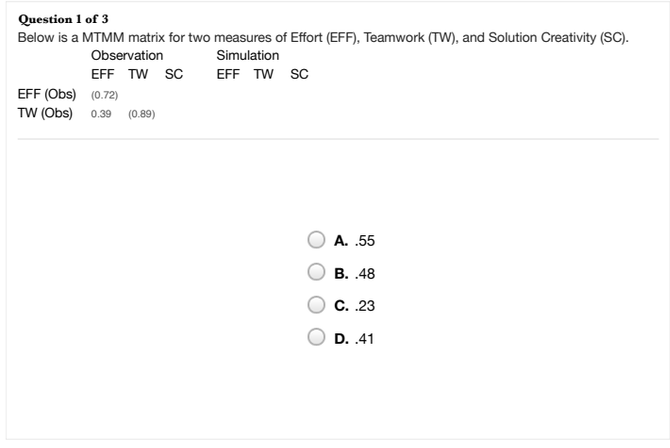
	Paper & Pencil		Observation	
	Dep	Sz	Dep	Sz
Depression (Paper & Pencil)	(.92)			
Schizophrenia (Paper & Pencil)	0.12	(.89)		

What we see above is some fantastic construct validity. (The fact that I made up the data doesn't hurt.) The convergent validity is great for the measures of depression (.72) as well as schizophrenia (.68). The discriminant validity coefficients are also fantastic, ranging from -.13 to .12. The other numbers on the matrix are reliability coefficients. It is common to place reliability coefficients in parentheses along the main diagonal. In the above matrix all of the reliability coefficients are good, with the lowest being .88 for our observational checklist of schizophrenia.

And now, because we can, how about a quiz (Review 1) to see if you can find your way around a MTMM matrix?

Problems with Construct Validity. There are a couple of problems with the construct validity process. The first is concerned with the second step in the convergent validity process. We need to find another test measuring the same construct against which we will compare our own test. How do we know that this other test actually measures the construct in question? We can't just take the word of the test developers on this issue. We need to examine its construct validity evidence. If this other test isn't a good measure of the construct in question, then it would be unwise for us to compare our test with it. (This is actually The Criterion Problem all over again.) The obvious solution is to always compare your test with the best available existing measure of the same construct.

REVIEW 1 Multi-Trait Multi-Method Matrix Quiz



The second problem is related to the first. It is more of a problem to the field of psychology than to any study in particular. How did someone establish that the existing test of the construct in question really measures that construct? They did their own convergent and discriminant validity studies. Which means that they compared their test to an existing test. And that test was compared with another existing test. Which in turn, was compared to another existing test. It's like a big circle. If we keep comparing our new tests to the old tests (and rejecting the new ones that fail to correlate), then all we'll ever have is more of the same old stuff. Well, what if I write a new test of intelligence that is different in conception and design than any of the existing tests? If my new test is really as different as I think, then it shouldn't correlate with the existing tests to a substantial degree. If I'm right about this new test, then there isn't anything available that should converge with it. A lot of tests are available for discriminant validity, but none for convergent. What to do? In such a case, I'll need to abandon the construct validity strategy and adopt a criterion-related strategy. I'll need to show that my test can predict meaningful, relevant external variables (e.g., school performance, job performance, etc.) better than the existing tests.

Item Response Theory

It's like waking up in the future.

Introduction

It's time for a major change. CTT represents basic measurement theory. But there are two newer measurement theories that can do things of which CTT could never dream. This chapter is about one of these newer theories, item response theory (IRT), a theory of measurement which dates to the work of Lord (1952). (We won't be discussing the other theory, Generalizability Theory.) To best explain what IRT can do for us, let's talk about the way things are with CTT.

With CTT a test has a single reliability coefficient, which of course leads to a single standard error of measurement (and SEE and SED and so forth). The reliability coefficient, standard error of measurement, and standard error of the difference all serve to convey, in different ways, the precision of our measurement. Better reliability means more precise measurement. And in CTT there is just one reliability per test and thus, just one standard error of measurement, which means that in CTT our test is equally precise for all test takers. If the standard error of measurement is 5.4, it's 5.4 for people with high scores, low scores, and scores near the middle of the distribution. According to CTT, that is. The reality may be different. Consider a test with a bunch of items average in difficulty and only a couple that are very easy and very hard. Given the limited number of items useful for measuring people of high and low ability, it would be darn near impossible for such a test to be equally precise for people of high, low, and average ability. IRT doesn't force such a limitation on us. With IRT, we may find that a given test is not equally precise for all ability levels. Thus, in IRT a test will not have a single standard error of measurement, rather the standard error of measurement for a test can be different for test scores of all levels. The typical pattern is one in which high and low scores have high standard errors of measurement (i.e., poor precision) and average scores

have low standard errors of measurement (i.e., good precision). Although this scenario is typical of many tests, it is not forced on a test with IRT. Of course, it is always possible (though unlikely) that a test could be equally precise for the entire range of test scores. If it is the case, then IRT will properly model precision. It is also possible that a test could be best (most precise) for people of high and low ability and worst for people of average ability (IRT will also model this situation properly), although this scenario is extremely unlikely.

So what does all of this measurement precision stuff mean to us? How will it affect our daily lives? Measurement precision is what allows us to be confident that two scores are actually different from each other. If you score an 89 on your science final and your friend scores an 88 on the same test, is this score difference likely due to just random error? Probably yes. You know that random errors may have raised your score or lowered your friend's score. If, however, our test is perfectly precise (SEM and SEE are both 0.0), we can be confident that a one point difference means your true score is actually higher; that is, the difference is not due to a random error. New example: Let's say you get a 90 and your friend gets a 70. Can we be confident that you actually know more? You might think, 20 point difference, that's pretty big, it can't be due to random error. But we need to know the standard error of the difference to answer this question. If our test has a huge standard error, then we can't be confident at all that your true score is greater. To sum all of this up, precise measurement is good. CTT models measurement precision in an unrealistic fashion. IRT models it in a realistic fashion.

Life with CTT also means that we have pretty much one way to score a test, called number-right scoring. With number-right scoring, a person's score is based on the number of items answered correctly. This number is often rescaled into some other number (e.g., percent correct, *z* scores, etc.),

but it all starts with the number of items answered correctly. IRT allows for another way to score a test, called maximum likelihood scoring. With maximum likelihood scoring, we attempt to determine what ability level best fits the pattern of right and wrong responses given to questions of known difficulty. As an example, let's say a person answers half of the easy questions correctly, but misses all of the items of average or high difficulty. In such a situation, we ask ourselves, is it likely that a person of high ability would perform in this way? No, it is not very likely (a high ability person would be unlikely to miss half the easy items as well as all of the harder items). Is it likely that a low ability person would perform in such a way? Yes, it is fairly likely. Thus, the score we give this person will be low. We'll discuss this more later.

Next up is a related topic. So related that it is pretty much the same topic. In CTT we never could properly deal with random errors (guessing, missing stuff you know) when we scored a person's response. Guess correctly and you benefit. Miss stuff you know, and you lose. But IRT's maximum likelihood scoring allows us to handle, to a limited extent, certain random errors of measurement which we'll call aberrant responses. That is, with maximum likelihood scoring these errors will have less of an impact on a person's score than with number-right scoring. Starting with our above example (half of easy questions answered correctly, everything else missed), let's say that our test taker also answers the hardest question on the test correctly. Number-right scoring gives him the extra point, but maximum likelihood scoring recognizes that a correct response on the hardest item (by someone who has missed half of the easy items in addition to pretty much every other item on the test) is an aberrant response. (Similarly, imagine a person who answers every question on the test correctly except for the easiest item. In such a case, this aberrant response would be ignored by maximum likelihood scoring.) Aber-

rant responses are not greatly considered in maximum likelihood scoring.

The Logic of Item Response Theory

IRT looks at the world through the lens of conditional probability. (That's the coolest sentence I've ever written.) With conditional probability the chances of something happening is dependent upon some other event. For example, what is the chance that you will win the lottery? Well that is dependent on whether you buy a lottery ticket. Your chance of winning is conditional upon whether you buy the ticket. If you do not play the lottery, your chance of winning is zero. But if you do buy a lottery ticket, your chance of winning is, um… well, still pretty much zero. Because it's the lottery. What we need is a better example. Your chances of running (specifically, finishing) a marathon is conditional upon whether you train. If you train hard before the marathon, you probably will finish the race. But if you do not train at all your chances are next to nil that you will finish.

Now let's apply this concept to testing. What are the chances that you will answer Item 1 on a test correctly? To answer this question, we need to understand the three components of IRT. A correct response to an item (first part) is conditional upon the test taker's ability (second part) and the characteristics of the item (third part). The part about test taker ability is easy to understand. If you are of high ability (you have a high standing on the measured construct), other things being equal, you will have a greater chance of answering the item correctly than a person of low ability. No guarantees here, but a better chance. Now let's focus on the item. Is Item 1 a hard item or an easy item? Is it a good measure of the construct in question or a poor one? Is it an easy item to guess correctly (maybe because it is poorly written) or very tough to guess? These issues are referred to as the item's characteristics. Let's just focus on the diffi-

culty one. If the item is easy and you are of high ability, what are your chances of answering the item correctly? Pretty good. If the item is hard and you are of low ability, what are your chances? Not good. That's the basis of IRT.

At this point you might have a few irritating thoughts bothering you like, "How do we know an item's characteristics?" Good question. The answer is that in order to know the item characteristics, we will have to do a study to analyze the items. This study will help determine how difficult the item is along with the other issues mentioned above. Once the item characteristics are known, we can use the items. The other question you might have is, "How do we know the person's ability?" A follow up question might be, "If we know the person's ability, why would we bother testing them?" You instincts are correct. We don't know a person's ability before we test them. But the nature of the IRT model allows us to turn things around to estimate a person's ability, which is pretty much the whole point of any test. Here's what I mean. When a person completes a test, what do we really know? We know the characteristics of the item, as established in a previous study. We also know which items the person answered correctly. That's what we know. We don't know the ability of the person, but based upon the two things we do know, we can estimate his or her ability. Example. Let's say we have 20 very difficult math questions. And let's say that our test taker answers 19 of them correctly. What are the chances that a low ability person could have answered 19 of 20 hard math questions correctly? Not good. What about a person of average ability? A little more likely, but still not much of a chance. What about a person of high ability? Ah, now we have something. It is very likely that a high ability person could answer 19 of 20 very hard math questions correctly. Thus, we estimate this person's ability to be high. The end. What I've just described is a good summary of how we score tests

in IRT. We'll come back to this scoring issue later in the chapter. I hope that this example illustrates that in IRT, it is all about conditional probability.

Definitions and More Definitions

Unfortunately, IRT uses what amounts to a whole new language. And there's no way around it. Many of these new terms will be three words long and, hence, will have a Three-Letter Acronym (TLA).

Theta: Our estimate for the ability of the test taker. When we score a test using IRT, the scores will be called theta (i.e., θ). Theta is typically scaled like a *z* score, so you know that a theta of +1.5 is a high score. Of course, we are free to rescale thetas into whatever metric we like (e.g., *T* scores), but we'll stick with thetas in *z* score units.

Homogeneous Sub Population *(HSP)*: A group of people who all have the same theta score on a given test. Imagine that we know every person's theta score and that we sort these people into groups with the same theta scores. We might have 83 people with a theta of -1.7. That's an HSP. We have another 94 people with a theta of -1.6. Boom. That's another HSP. Because the main estimate of IRT is unidimensionality, we know the people within a given HSP have something in common: the same estimated ability or standing on the construct in question. People from different HSPs are different in one important way. They have different standings on the constructs. The difference might be small (e.g., -1.7 vs. -1.6) or it might be big (e.g., -1.7 vs. +1.9), but there is a difference.

Probability of a Correct Response *(PCR)*: The percent of people from a given HSP answering a given item correctly. To illustrate this concept, just imagine that we give Item 1 to the -1.7 HSP.

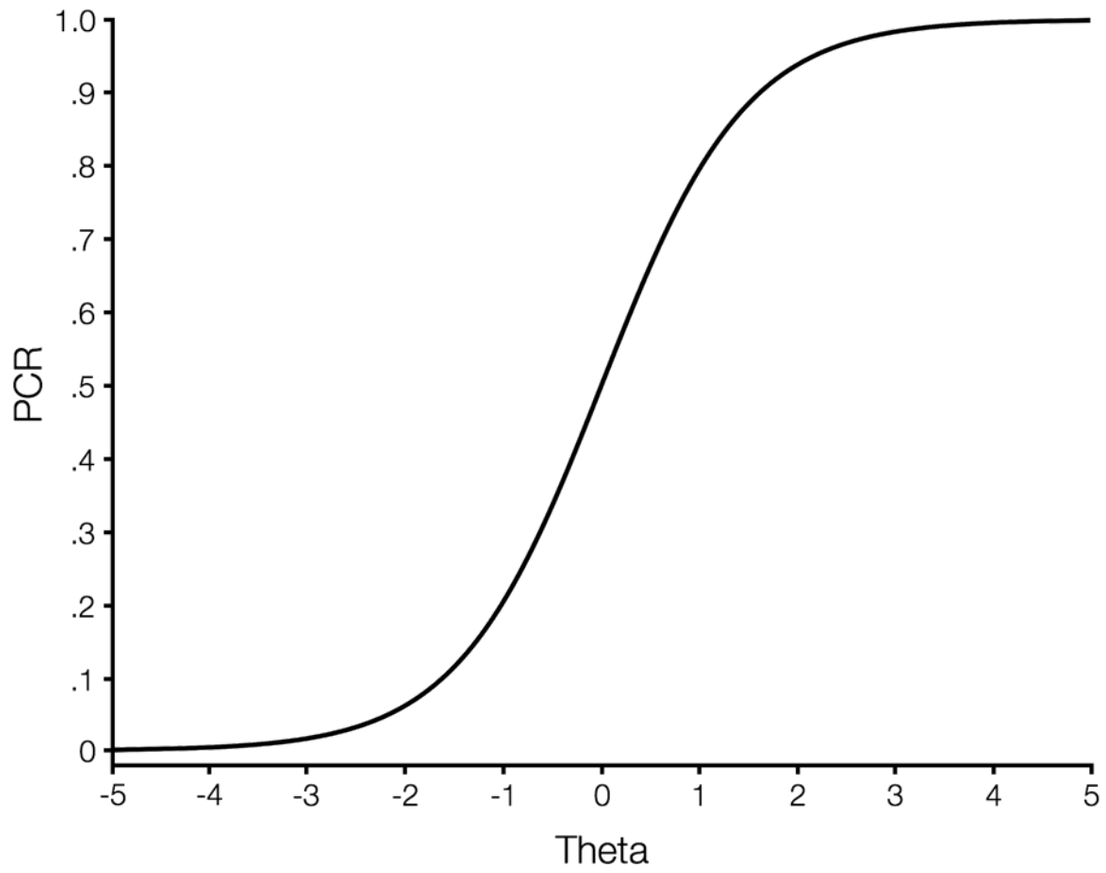
Then we compute the percent in that HSP answering Item 1 correctly. If 4 of 83 people answer it correctly, then the PCR for that HSP is .048 (or 4.8%). Thus, each HSP will have a PCR for each item. Which leads to…

Item Characteristic Curve *(ICC)*: A graph of PCRs for all the different HSPs. Naturally, there will be one ICC per item. I wasn't lying about that whole new language thing.

An example of an ICC for a hypothetical item (we'll call it Item Gamma) is shown in Figure 1. Notice how you can relate PCR to a given theta (e.g., PCR of .5 is associated with a theta of 0.0). Or you can start from theta and go to PCR (e.g., a group of people with a theta of -2 would be expected to get this item correct 5% of the time).

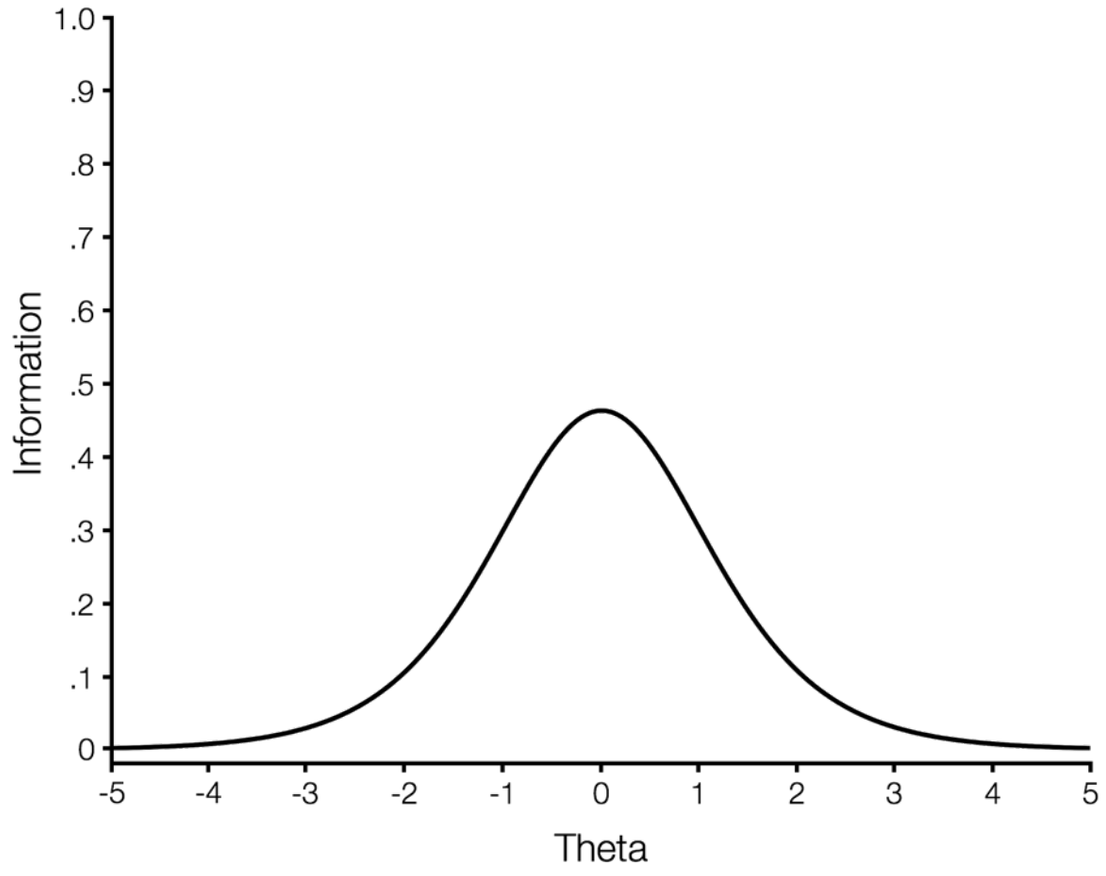
Information: The absence of uncertainty. If you're like me, you hate it when a term is defined as what it's not. So here's something better: measurement precision. Each item will contribute information to the test and the amount of information contributed will vary by theta.

FIGURE 1 Item Characteristic Curve



Item Information Function *(IIF)*: A graph of information at various thetas for a given item. Figure 2 is the item information function for the aforementioned Item Gamma. For this item (Item Gamma), information is maximized for theta

FIGURE 2 Item Information Function



scores of around 0. This item offers almost no information for people with scores beyond +2 or -2.

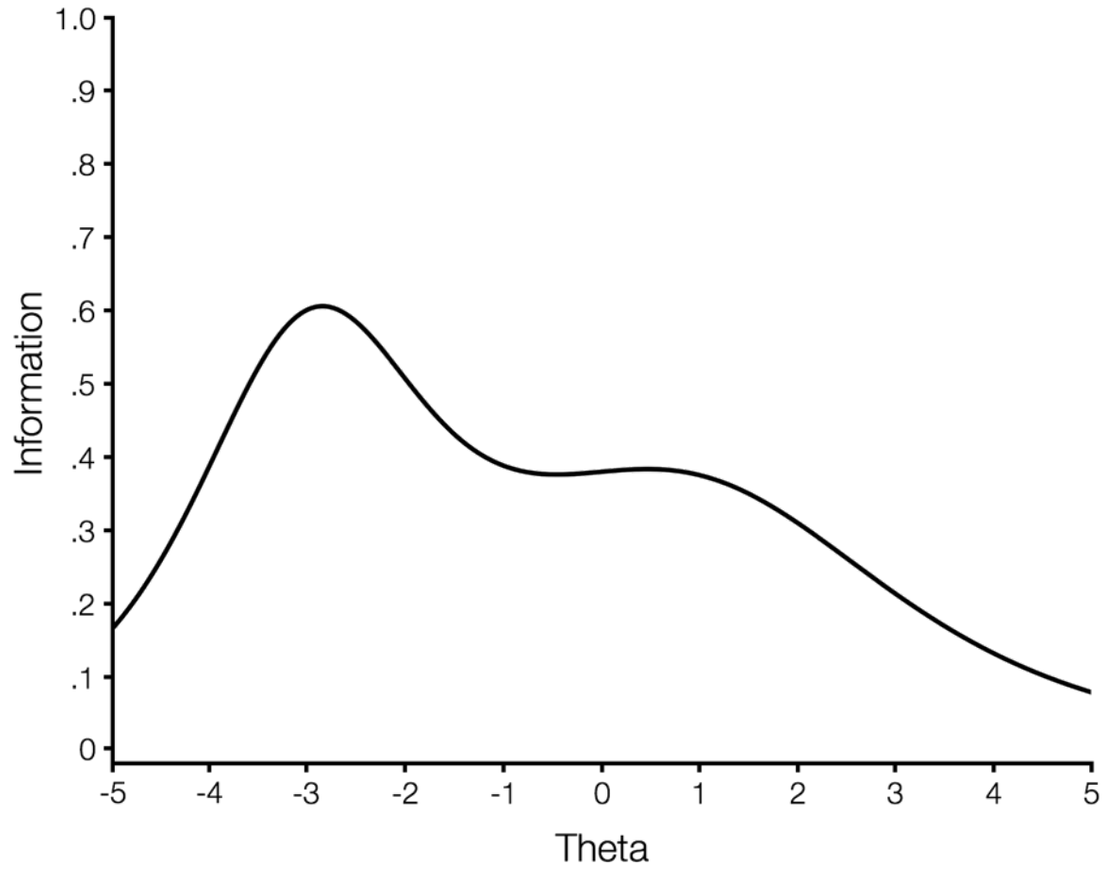
If you compare the ICC (Figure 1) to the IIF (Figure 2) for our Item Gamma (Why Gamma? Because Greek letters are cool.), you can see that the ICC is essentially flat at the extremes and has its sharpest slope at around 0 theta. Thus, we can correctly infer that the slope of the ICC is a prime determinant of information. Sharper slopes mean more information. We'll come back to this point later, but we can also note that it will be impossible to construct a single item that yields a ton of information for all thetas. Why? Although an item can have a gradual slope that spans all of the theta range, it cannot have a sharp slope for more than a small portion of the theta range. If an item has a sharp slope, the steep slope is limited to just one area. Thus, a single item can only yield a large amount of information at one spot. That said, what if we put together a 100-item test, each item with a sharp slope but at different spots along the theta range? Our test as a whole would have a lot of information across the theta range. Which leads to our next concept...

Test Information Function *(TIF)*: The sum of the IIFs for all of the items on the test. If we have a five-item test, each item has an IIF. But we can sum the information offered by each item at a

given theta (e.g., Item 1 yields .29 units of information at -2 theta, Item 2 yields .05 units of information at -2 theta, etc.) to yield an information value at that theta for the entire test. Repeat for other theta values. This resulting TIF will tell us how much information the test as a whole (i.e., collection of items) offers.

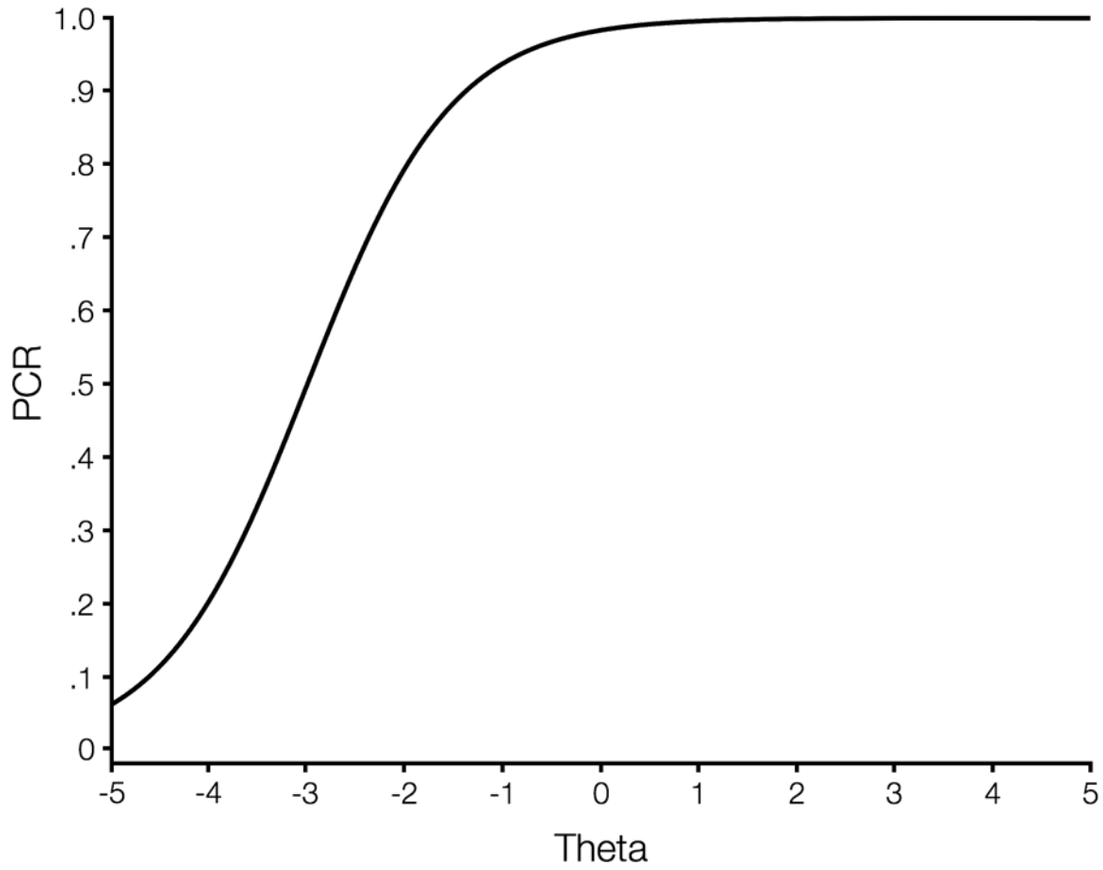
Figure 3 is a TIF for a five-item test of unknown origin (we'll call it Test Epsilon). As you can see the test offers at least .3 units of information for thetas ranging from -4 to +2. As you can also see, this test is most precise (i.e., offers the most information) for thetas near -3, but also offers a non-trivial amount of information between -4 and -2. The information offered between -4 and -2 is due to one of the items (Item 1) that functioned very well in that range. Figure 4 is the ICC for that item. Note how all of its slope is between -4 and -2, meaning that all of its information is offered in the -4 to -2 area.

FIGURE 3 Test Information Function for Test Epsilon



At this point, you might be wondering just how much information is a good amount of information. More is better, but the upside of information is unbounded. Yes, I know the *y*-axis of the graph stops at 1.0 but that was just a choice I made for convenience. Information ranges from

FIGURE 4 ICC for Item 1 (b = -3.0) from Test Epsilon



zero to positive infinity. Thus, information is unstandardized. Which leads to…

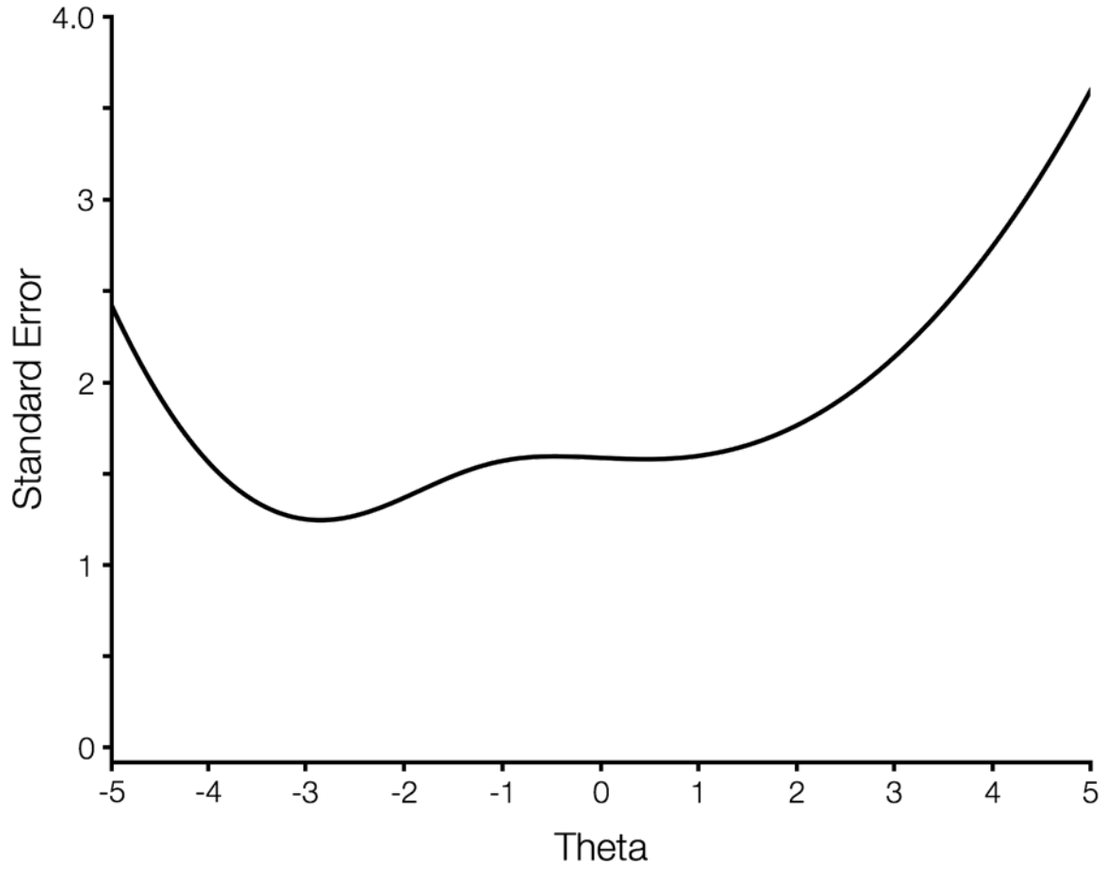
Test Standard Error Function *(TSE)*. For some reason the word *function* doesn't get to be a part of the acronym. Probably due to some rule about TLAs. What is the TSE Function? It is a conversion of the TIF into a standard error metric. Best of all, it's a simple equation. First, just determine how much information a test offers at a given theta. Using our TIF from above, our five-item test offers approximately .4 units of information at a theta of -1. I hope you successfully found that on the graph (make sure to look at the TIF). Then, plug into the following equation.

$$SE_\theta = \frac{1}{\sqrt{Info_\theta}}$$

Where:
SE_θ is the standard error of an item at θ.
$Info_\theta$ is the information by provided by an item at θ.

Thus, the standard error for our hypothetical five-item test at a theta of -1 is 1.6. Given that a standard error of 1.0 equates to a full standard deviation, a standard error of 1.6 is pretty bad. Poor measurement precision. A graph of the TSE function for our five-item Test Epsilon is shown in Figure 5. If you compare it with the TIF shown in Fig-

FIGURE 5 Test Standard Error Curve for Test Epsilon



ure 3, you can see that one is just the inverse of the other.

Finally, you might wonder if there is a Test Characteristic Curve or an Item Standard Error Function. There are, but they are not as useful as what we have listed above.

The Incredible World of IRT

Now that we have our terminology clear, we can understand more of the IRT world. Let's play "What if?" What if we had an item (it needs a cool name – we'll call it Item Foxtrot) with an ICC as shown in Figure 6? Now let's look carefully at Item Foxtrot. For people with thetas less than 0, what are their chances of answering this item correctly? Zero. Not even a slight chance. Not even by guessing (it should be obvious that we'll never see this with real data). For people with thetas greater than 0, what are their chances of answering this item correctly? 1.0 (i.e., 100%). No one will miss it, not even by some sort of random error. (Don't try to figure out what will happen if a person's theta exactly equals zero. You'll develop a serious headache.) Now let's turn it around. If I give this item to a person and they get it right, what do I know about his or her theta? It is definitely greater than zero. It may be 0.1 or it may be 2.8. I can't say how much higher, but it's definitely

FIGURE 6 Item Characteristic Curve for Item Foxtrot

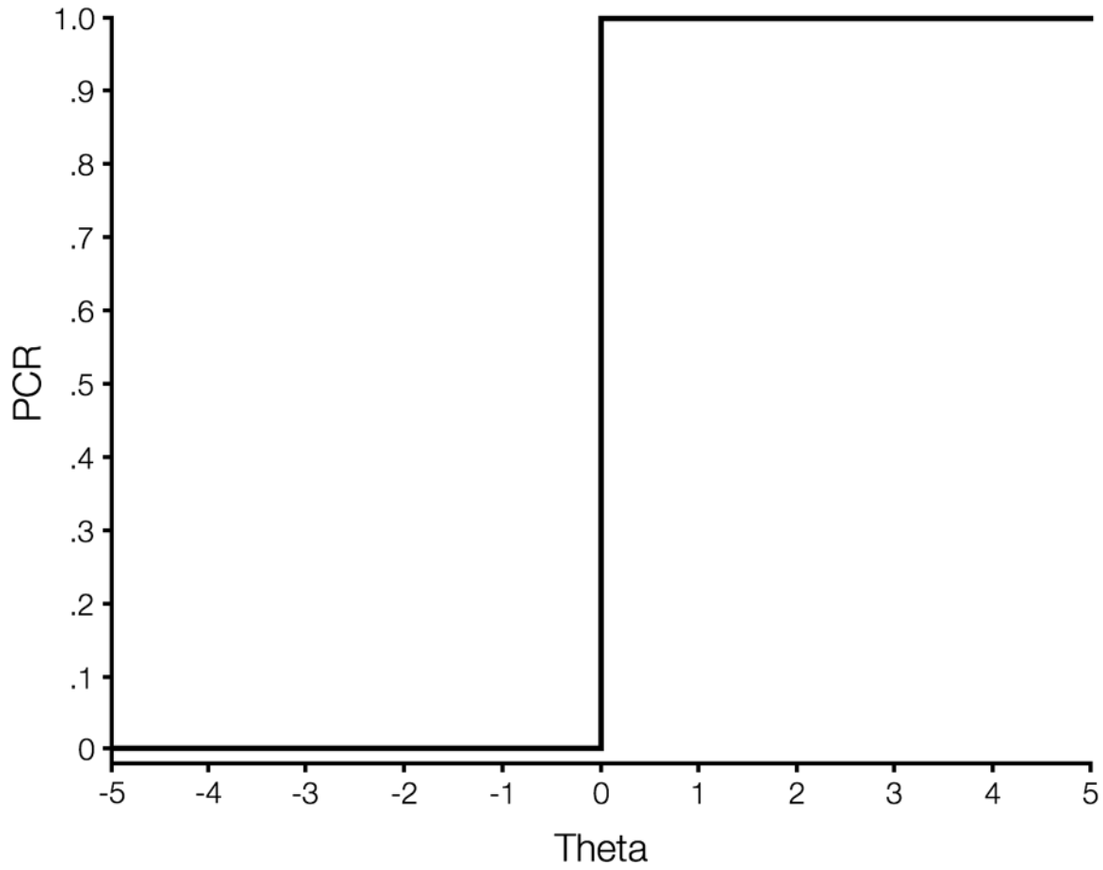
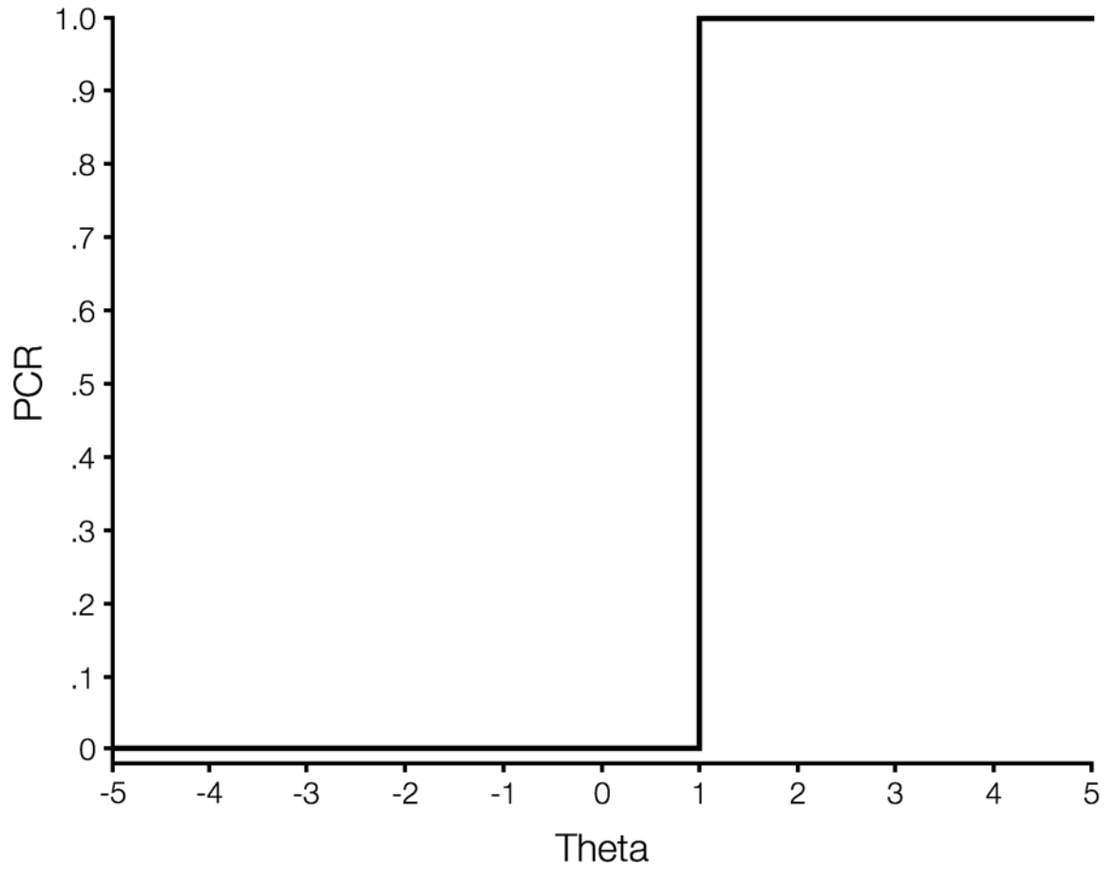


FIGURE 7 Item Characteristic Curve for Item Tango



greater than zero. Similarly, if someone misses this item, then we can say with complete confidence that his theta is less than zero. We don't know how far below, but it is less than zero.

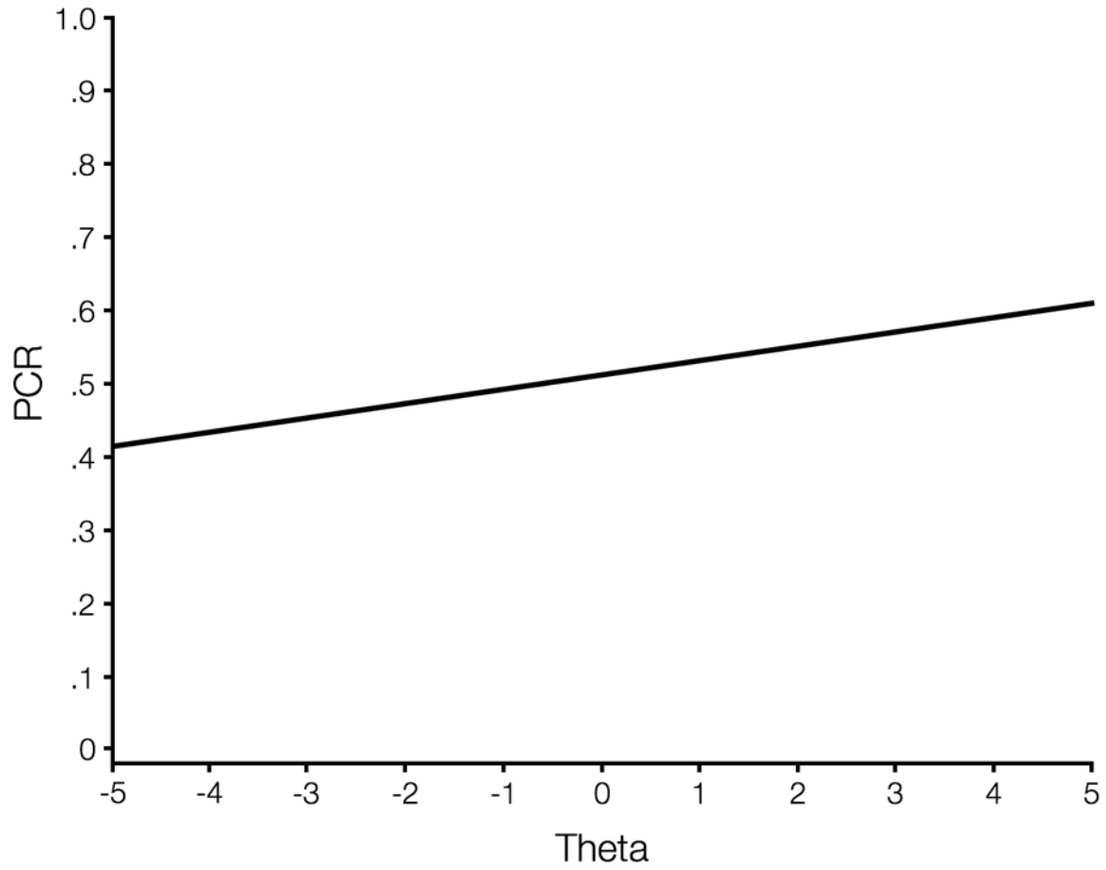
At this point you might be saying, big deal. But this is only one item. Let's say we have a two-item test with the second item (another cool name needed, let's see... Item Tango) presented in Figure 7. Now imagine that a person answers Item Foxtrot correctly (meaning that his theta is greater than 0.0) but misses Item Tango (meaning that his theta is less than 1.0). We can conclude with absolute confidence that this person's theta is be-

tween 0 and 1. With more items, I can be more precise.

Now at this point you might be thinking, "OK, smart guy, what if the test taker misses Item Foxtrot and gets Item Tango correct?" Your mistake is that you're thinking of the real world where people can guess correctly and make all other sorts of random errors. Items Foxtrot and Tango exist in a fantasy world where that doesn't happen. It can't happen with our two items. Check the ICCs – if a person has an ability greater than 0, they must answer Item Foxtrot correctly.

What was the point of all of this? It was to show you how the item characteristics, the response of the test taker, and scoring a test in IRT all relate to each other. As a point of contrast to our earlier super items (Foxtrot and Tango), consider the item presented in Figure 8. Let's say someone answers the Figure 8 item correctly. What can we conclude about this person? Not much. Look at people with a theta of -3, which is very low ability. These people have roughly a 45% chance of answering this item correctly (PCR = .45). Compare that with people at +3. PCR there is around .55 (or 55%). Thus, the chance of answering this item correctly for someone very low in ability as compared to someone very high in

FIGURE 8 Yet Another Item Characteristic Curve



ability increases by only ten percentage points. Thus, if you answer this item correctly, I can't conclude much of anything about your ability with any confidence. You might be low in ability. You might be average. You might be high. I just don't know. So to summarize, what we see here is a worthless item. Its IIF would be nearly flat and close to zero for all theta scores.

What have we learned by this exercise? In IRT, we like items that have a sharp slope at some point along the theta range. We want a lot of these items and we want them to have their sharp slopes at different points, so that our test offers information (remember sharper slope means more information) at every location along the theta scale that we care about.

The Ugly Details

Let's talk about where this ICC actually comes from. In the *Definitions* section, I described the process as one in which the item is given to people from different HSPs and plot the PCR for each group (set aside the issue of how we know each person's ability). We could do that. Doing so is called *empirical IRT*. But, it's a little more efficient if we model the shape of the ICC with a smooth function that we could describe with just three terms. That is, I can graph any ICC correctly if I know just three numbers for that item. This is called the three-parameter model of IRT for obvious reasons. There are one- and two-parameter models, but they are clearly excluding something important.

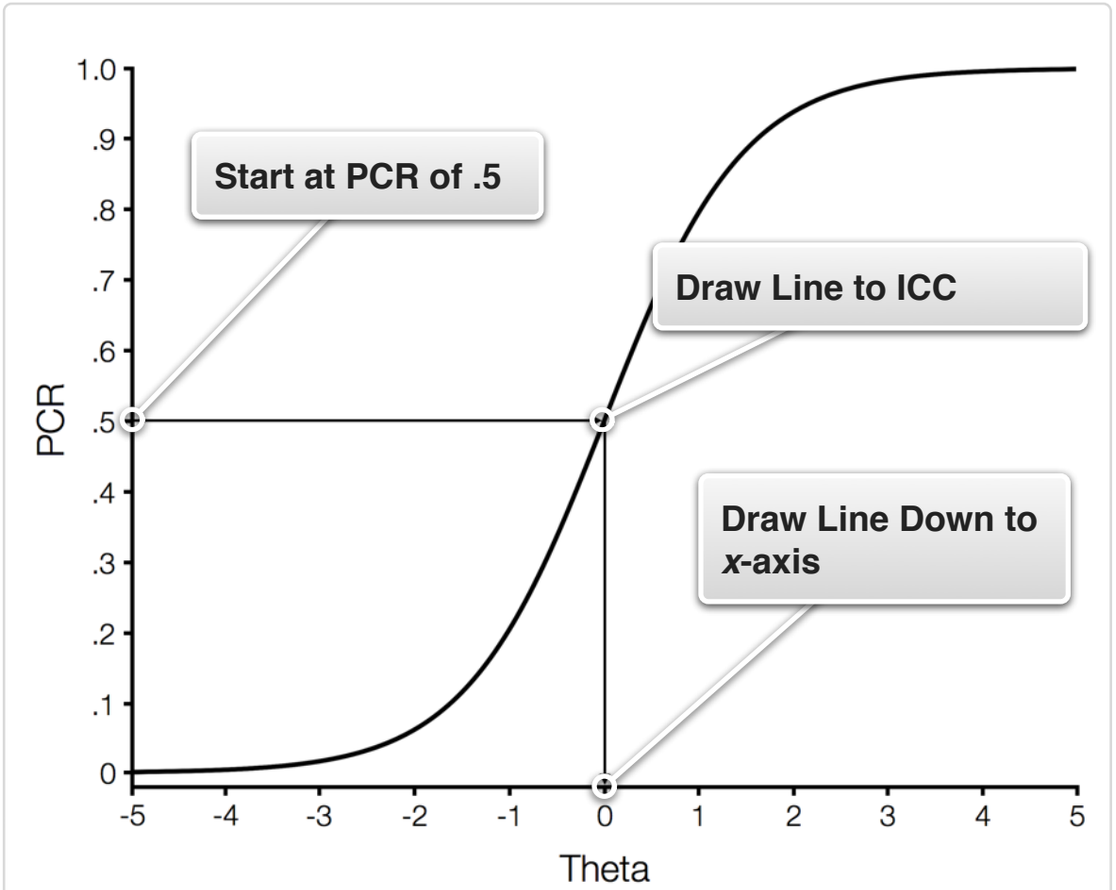
The three parameters are: difficulty (called the *b* parameter), discrimination (the *a* parameter), and lower-asymptote (*c* parameter). Note how the letters are not in alphabetical order. A normal person would have labeled them *a*, *b*, and *c*, but obviously IRT was not developed by a normal person. I'm sure there is a long and interesting story about this, but I don't want to hear it.

Difficulty *(b)*: Describes how difficult the item is. It is analogous to *p* value from the CTT days. The difference is we will describe difficulty in IRT as the ability level required to have a 50% chance of answering the item correctly. It is very easy to determine the difficulty of the item by looking at the ICC. Just start at a PCR of .50 (50% chance of success) and draw a line to the ICC, then come straight down and draw a line to the theta.

An example has been diagramed in Figure 9. In the case of this item, the value of *b* is 0.0. This item is average in difficulty. Why? A person must be average in ability to have a 50% chance of answering this item correctly.

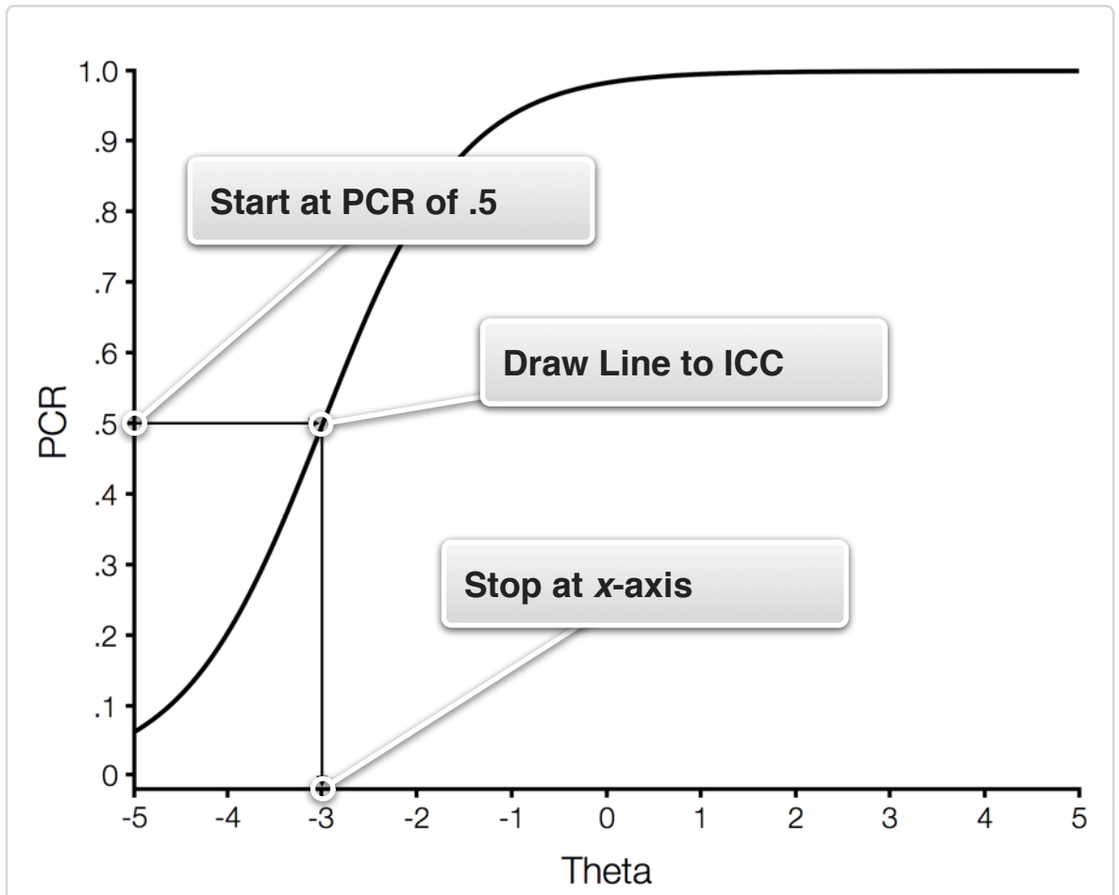
A new item is shown in Figure 10. This item is easy. It has a *b-paramter* of -3.0. That means a person of very low ability (-3.0 is very low) has a 50% chance of answering this item correctly. As a point of comparison, what is the PCR associated with a theta of 0? I'm just eyeballing it here, but it

FIGURE 9 Item Difficulty (*b* Parameter) Illustrated



looks to be about .97 (or 97%). Now that's much better than the previous item. Thus, easier item. So, to summarize, if *b* is close to zero, the item is average in difficulty. If *b* is seriously negative, the item is easy. If *b* is strongly positive, the item is

FIGURE 10 Item Difficulty (*b* Parameter) Illustrated – Yet Again

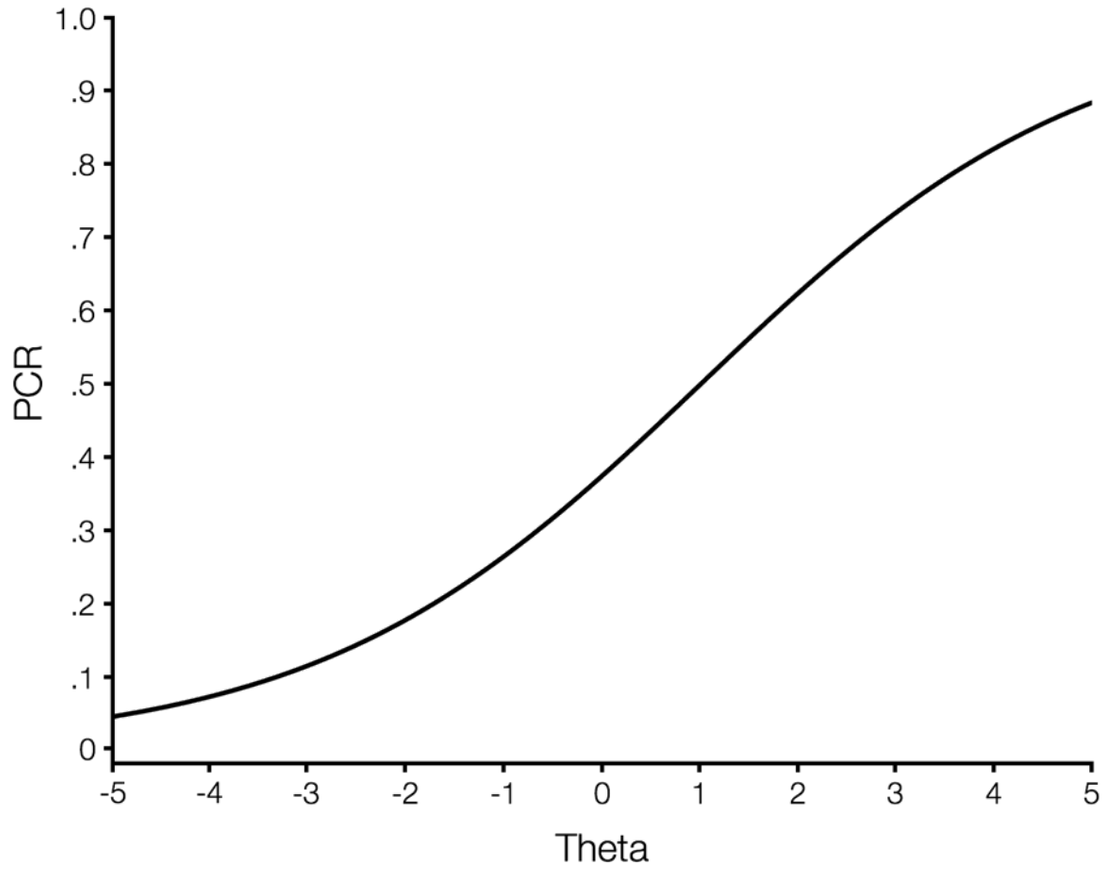


very difficult. There, now we've done the first parameter. Moving on.

Discrimination *(a)*. Item discrimination describes the relationship between a correct response to an item and test taker ability. We've actually discussed this before; we just didn't know it. Do you remember those stair-stepped shaped ICCs from before (Figure 6 and Figure 7)? Those kinds of items do a perfect job at determining whether a test taker has enough ability to answer the item correctly. In fact, they do such a great job, they are not realistic items. Totally fictional. Real items don't work that well. Now examine the item that had an almost flat ICC (Figure 8). As we discussed, it was terrible at relating test taker responses to test taker ability. What we have been discussing this whole time was the item's discrimination, or *a*, parameter. Sharper slopes mean better discrimination. Sharper slopes yield more information, which means less standard error, which ultimately means better measurement precision. In short, we like items with high *a* parameters. *a* can range from zero (not good) to positive infinity. To compute *a*, we need to compute the slope of a line at a point on the curve in which the slope is at its strongest, which is at or near the location of *b*. We

can compute the slope at any point on the curve. Only one point, however, will have the maximum slope. This point happens to be at or very near the theta associated with a PCR of .5. Computation of slope at a point on a curve requires differential calculus. I don't know about you, but I'd rather not deal with that. We'll just let computers do the dirty work for us. That said, it's very easy to visually compare two or more ICCs and determine which has the better *a* parameter. ICCs for two items are presented in Figure 11 and Figure 12. It is no challenge to see that the ICC in Figure 12 has the steeper slope, and thus, the higher *a* parameter. The *a* parameter for the item in Figure 11 is .3, whereas the *a* for the Figure 12 item is .8. As mentioned, there's no way to easily compute those numbers from looking at the graph; I used a computer to obtain them. Notice that both items have the same *b* parameter (+1.0). They are equal in difficulty, but the for Figure 12, a correct response is more closely linked to test taker ability.

FIGURE 11 Item Discrimination (*a* = .3)



Lower Asymptote *(c)*. The *c* parameter, also called the *pseudo guessing* parameter, tells us the percent of very low ability people who answer the item correctly. It is analogous to the *y*-intercept from regression days. To estimate *c* from the ICC, if the ICC has flattened out (the *asymptote* part), simply find the point where the ICC intersects the

FIGURE 12 Item Discrimination (*a* = .8)

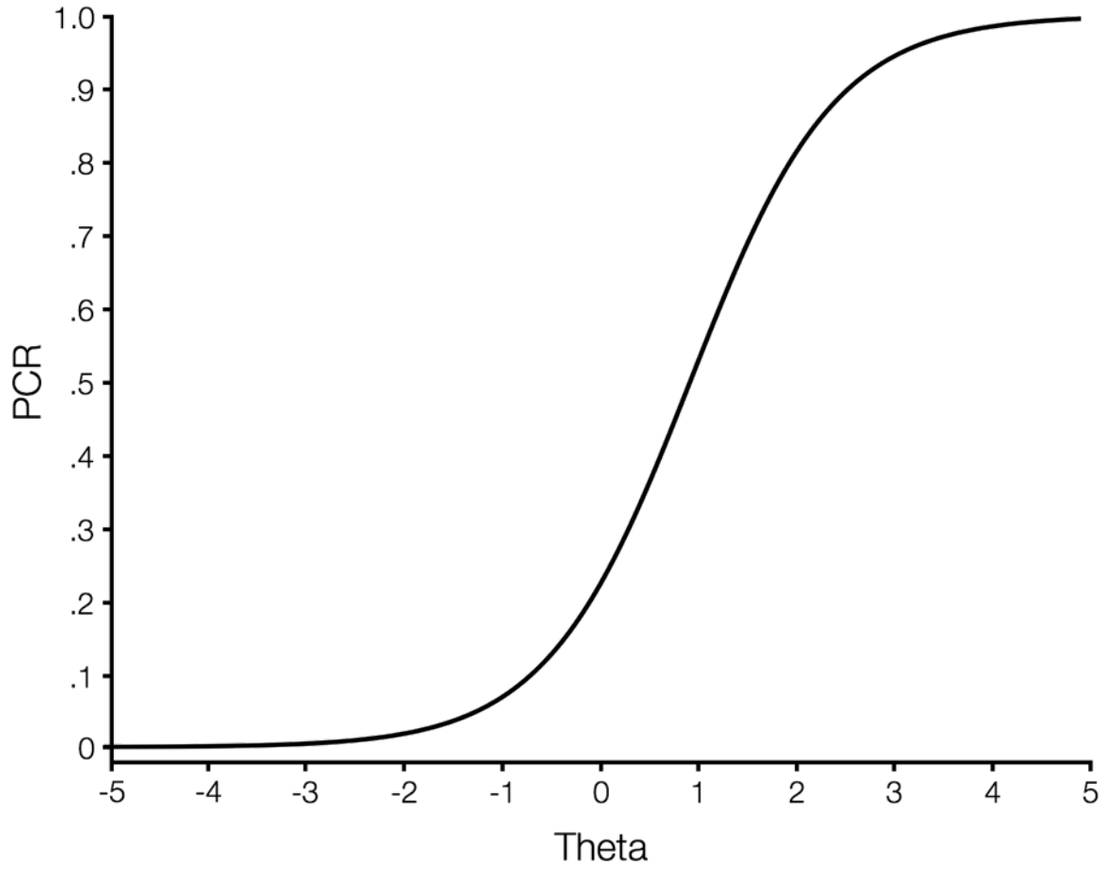
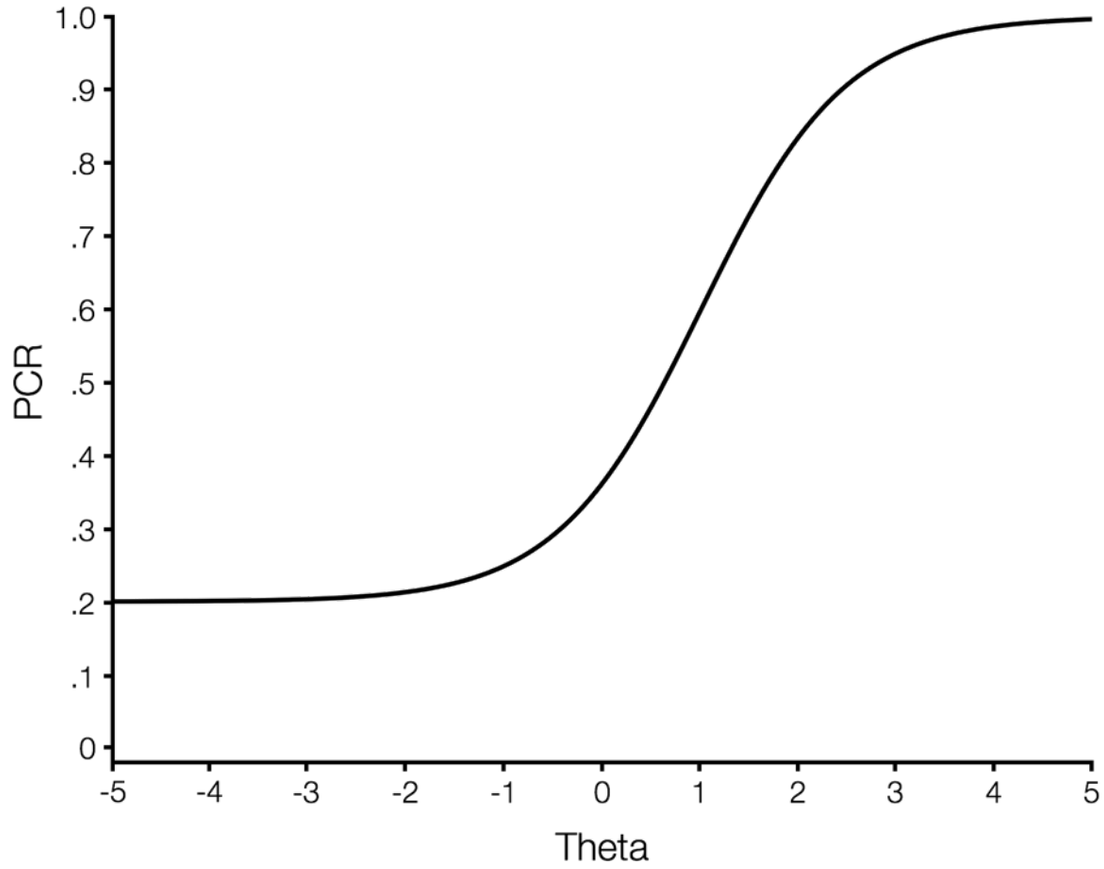


FIGURE 13 Lower Asymptote (*c* = .2)



y-axis. To understand the *c* parameter, let us once again examine Figure 12. As you can see, at the low end of the theta range, the ICC is flat and intersects the *y*-axis at 0.0. Thus, the *c* parameter is zero, meaning that zero percent of very low ability test takers answer the item correctly. (Given that people have a bad habit of guessing correctly in the real world, *c* will not be zero for real ICCs.) Contrast Figure 12 with Figure 13. Figure 13 has the same *b* (+1.0) and the same *a* (0.8) as Figure 12, but it has a *c* of .20. This means that people of very low ability (i.e., beyond -2.0) still have a 20% chance of answering the item correctly. We don't know why. Probably a guessing issue. (By the by, if

you're wondering why the *b* parameter in the second ICC looks like it's less than 1.0, that happens when *c* is greater than 0.0. Non-zero *c* parameters actually shift the graphical representation of the *b* parameter a little. Don't worry about it, though. It's not a big issue.)

More about *c*. Non-zero *c* parameters are bad. *c* is the antitheses of *a*. *c* actually reduces information, and thus, measurement precision. I have seen many items with great *a* parameters that also had large *c* parameters and, thus, were worthless. To summarize, we like items with large *a* parameters, low *c* parameters, and probably a variety of *b* parameters.

PCR Is as Easy as b, a, c. An item's characteristics can be described with just three numbers: *b*, *a*, and *c*. Everything we need to know about the item is captured by them. Better yet, we can use these three numbers to draw the ICC and make exact computations of PCR for a given theta. To do so, we need to know the equation that relates theta to PCR. And here it is.

$$PCR_\theta = c + (1 - c)\frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

Where:
PCR_θ is the probability of a correct response at a given θ.

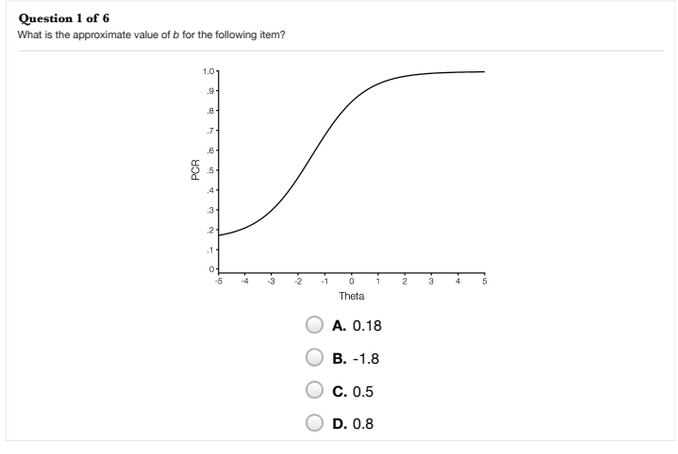
First thing you may be thinking is, wait, what is *D*? And where did that *e* come from? *D* is just a constant with the value of 1.702. *e* is an arithmetic function and is the inverse of the natural logarithm (you may have noticed an *e* button on your calculator). The only variables in the equation are the familiar *b*, *a*, and *c*. We plug in *b*, *a*, and *c* for a given item and compute for thetas ranging from -5 to 5. Now we have exact PCRs for every theta, allowing us to generate our graphs or perform exact calculations.

What about the IIF, you say? How do we draw that graph, you say? Well, the IIF is just another equation with a bunch of familiar parts.

$$Info_\theta = \frac{D^2a^2(1-c)}{[c + e^{Da(\theta-b)}][1 + e^{-Da(\theta-b)}]^2}$$

Once again, just plug in *b*, *a*, and *c* (*D* still equals 1.702) and compute information across a range of thetas. Graph if desired. If you like playing around with equations, you can see how raising *a* increases information. You can also see how raising *c* lowers information.

REVIEW 1 Item Characteristic Curves



Estimating Theta

This is it. This is where we can obtain one of the big benefits of IRT. We will now find out how to score a test using maximum likelihood scoring. No more number-right scoring for us. To explain maximum likelihood scoring, recall that we said there are three components to IRT: a correct response to an item (first part) is conditional upon the test taker's ability (second part) and the item characteristics (third part). To put these three components together, I'll quote myself from earlier in the chapter.

> When a person completes a test, what do we really know? We know the characteristics of the item, as established in a previous study. We also know which items the person answered correctly. That's what we know. We don't know the ability of the person, but based upon the two things we do know, we can estimate his or her ability. Example. Let's

say we have 20 very difficult math questions. And let's say that our test taker answers 19 of them correctly. What are the chances that a low ability person could have answered 19 of 20 hard math questions correctly? Not good. What about a person of average ability? A little more likely, but still not much of a chance. What about a person of high ability? Ah, now we have something. It is very likely that a high ability person could answer 19 of 20 very hard math questions correctly. Thus, we estimate this person's ability to be high.

The basic premise is a simple one. Identify which theta best fits this person's pattern of responses given the item's characteristics (*b*, *a*, and *c* parameters). The math is not all that bad, but it is a little tedious. There are a few options as far as equations go. I'll give you the simple one, which I call, the simple one.

$$LH_\theta = \Pi[PCR_\theta(SR) + (1 - PCR_\theta)(1 - SR)]$$

Where:
LH_θ is the likelihood (or probability) that someone with a given theta could make correct responses to the items that the test taker answered correctly (and the converse).
SR is scored response to the item (1 if correct, 0 if incorrect).

The capital *pi* symbol (Π) at the front means that this is a sequential operation, but instead of summing (as with Σ), we multiply. Notice that the plus sign divides the equation into two halves. One of these halves will be zero every time due to the *SR* component, which is always 1 or 0 (if it is 1 for the left side, then it's 0 for the right – and the converse). So, we're always dealing with just half of this equation every time, the left half if the response was correct and the right half if the response was incorrect.

We will use this equation across items for a given theta for the test taker, yielding a probability that that a person with this theta could have made these responses. Before we get to the details, let take a side trip down a road called joint probability.

If you roll a single die, what is the probability that you will roll a six? The answer is 1/6. If you roll a die three times, what is the probability that you will roll three sixes? The answer is 1/6 x 1/6 x 1/6, which equals 1/216. We multiply the probabilities of the individual events to determine the probability that all of these events will occur (this multiplication is what is conveyed by the capital *pi* symbol, Π, in the maximum likelihood equation). Let's change up the dice example just a little bit. Still rolling the die three times, but now let's ask, What's the probability that the first two rolls will be sixes and the last roll will not be a six (that is, 1, 2, 3, 4, or 5)? To figure out that last part, we need to know the probability of not throwing a six. The probability of something not happening is one minus the probability of it happening. The probability of not throwing a six is one minus the probability of throwing a six. Thus, it's 1 - 1/6, which equals 5/6. So for our new version of the dice example, the probability of throwing sixes on the first two rolls and not throwing a six on the last roll is 1/6 x 1/6 x 5/6 = 5/216.

Maximum likelihood theta estimation operates in the same manner. PCR tells us the probability of a correct response to an item for a person of a given theta. What if the person missed the item? We use 1 - PCR. Now let's apply our last dice example (two sixes followed by a not six) to a three-item test. What's the probability that a person of a given theta will answer the first two items correctly and miss the last item? Like the dice example, we need to find the probability for each event and then multiply to obtain the joint probability. So it's PCR at the given theta for Item 1, PCR at the given theta for Item 2, and 1 - PCR at the

given theta for Item 3. Then, we multiply. This joint probability tells us the probability that a person with a given theta will answer the first two items correctly and miss the third. Just like the dice example. This process is what occurs with the maximum likelihood theta estimation equation.

That wasn't so bad. The kicker is that for maximum likelihood theta estimation we have to repeat this process across a range of theta scores. We only had to do this dice thing once, but we have to ask ourselves what the probability is that a test taker of a given theta could answer Items 1 and 2 correctly while missing Item 3 *for all possible thetas*. The theta with the greatest probability value is the most likely theta for our test taker, hence the name *maximum likelihood estimation*. In a sense, we're asking a series of probability questions (How likely is it that a person with a theta of -3 could have answered these questions correctly? How likely is it that a person with a theta of -2 could have answered these questions correctly? And so on.) and are picking the answer with the greatest probability value.

An example for our three-item test with a test taker who answered Items 1 and 2 correctly and Item 3 incorrectly is offered in Interactive 1.

Based on our calculations, the most likely theta for our test taker is +1.0 (actually, +.61 if we use the refined estimate). Let's step back and ask ourselves, Does this estimate make sense? Does it make sense that a person who answered

INTERACTIVE 1 Maximum Likelihood Theta Estimation

MAXIMUM LIKELIHOOD
THETA ESTIMATION

two items correctly, but missed the most difficult item would have a theta of +.61? Well, I guess so. It sure makes more sense than a theta of -1 (If it's that low, how did the test taker answer two items correctly?) or +2.5 (If it's that high, how did the test taker miss the third item?). An ability estimate of just above average average makes the most sense out of the test performance.

Finally, you may recall that I mentioned at the beginning of the chapter that maximum likelihood estimation allows us to reduce the impact of guessing and other random errors. An aberrant response (e.g., missing the easiest question while answering all of the hardest questions correctly) will be ignored with maximum likelihood scoring given a large number of test items. So we have that going for us, which is nice.

Computer Adaptive Testing

12

No one is allergic to this CAT.

Introduction

Now we'll take everything we learned in Chapter 11 and put it to good use. This chapter is all about computer adaptive testing (or CAT). Anyone can take a test and put it on a computer. That isn't a challenge. The hard part is to make it adaptive. An adaptive test is tailored to the ability of the person taking the test. Recall that in Chapter 9 we discussed difficulty-based item analysis. In difficulty-based item analysis a good item is one that is matched to the ability of the test taker. Well then, a bad item must be one that is poorly matched to the ability of the test taker (e.g., giving an easy item to a high ability test taker). What's so bad about that? It's bad because the test taker's response to that item doesn't tell us much about the test taker. A high ability test taker will likely answer an easy item correctly. If he should happen to miss an easy item, it's likely due to random error. Thus, either outcome isn't very informative. We would be a lot better off if we had given this high ability test taker a hard item. Now that would tell us something about the test taker.

Thus, we want to administer only items with difficulties that are a good match to the ability of the test taker. The only problem is: How do we know the ability of the test taker? (To be pedantic, if we knew the person's ability, we wouldn't need to give them the test.) Until now, the only way out of this conundrum were those situations where we had a general idea about the test taker's ability. For example, if we are giving a math test to a group of first grade students, the ability level for every one of those test takers will be low compared to the general population. In such a case, we want to use a lot of easy items on our test. But even that system is largely inefficient. Moreover, there are many other times when we don't have the first clue as to our test taker's ability. What then?

Well, we could give our test takers short pre-tests. That is, we give everyone a ten question pre-test, score it, and then give the test takers one of three versions of the main test (a hard, medium, or easy form) based on their pre-test score. The pre-test serves as our estimate of ability for the rest of the test, and we pick a set of items to best match that ability estimate. Any problems with the pre-test (e.g., missing items that you know) means that a test taker was stuck with the wrong version of the main test. Not good. With computers, we can change, or update, our estimate of a person's ability after every question. Thus, let's say our Mr. Test Taker* makes a few unpleasant random errors in the first few items, missing stuff he knows. Because he's missed a number of items, we don't think much of his ability at this point. But Mr. Test Taker starts paying attention and makes correct responses to the items he should get correct. We can now update our estimate of his ability. Based on this new estimate of his ability, we can give him the hard questions he deserves. Thus, computers allow us to update our estimate of a person's ability after every question, further allowing us to pick from our pool of items the best item (defined in terms of the match between item difficulty and test taker ability) for the test taker. That's adaptive testing.

*Not his real name

Now this whole adaptive testing thing flies in the face of one of the basic rules of measurement: standardize testing conditions (including test content) as much as possible. If we give two people two different versions of a test, how can I compare their scores? Wait, you say, don't the ACT people do that very thing all of the time? You bet they do. The items you get in October are not going to be the same items that your best friend gets when he takes the test in November. So they are violating this most basic of rules. We are generally not troubled by this issue because the ACT people spend a great deal of effort making sure that test given in

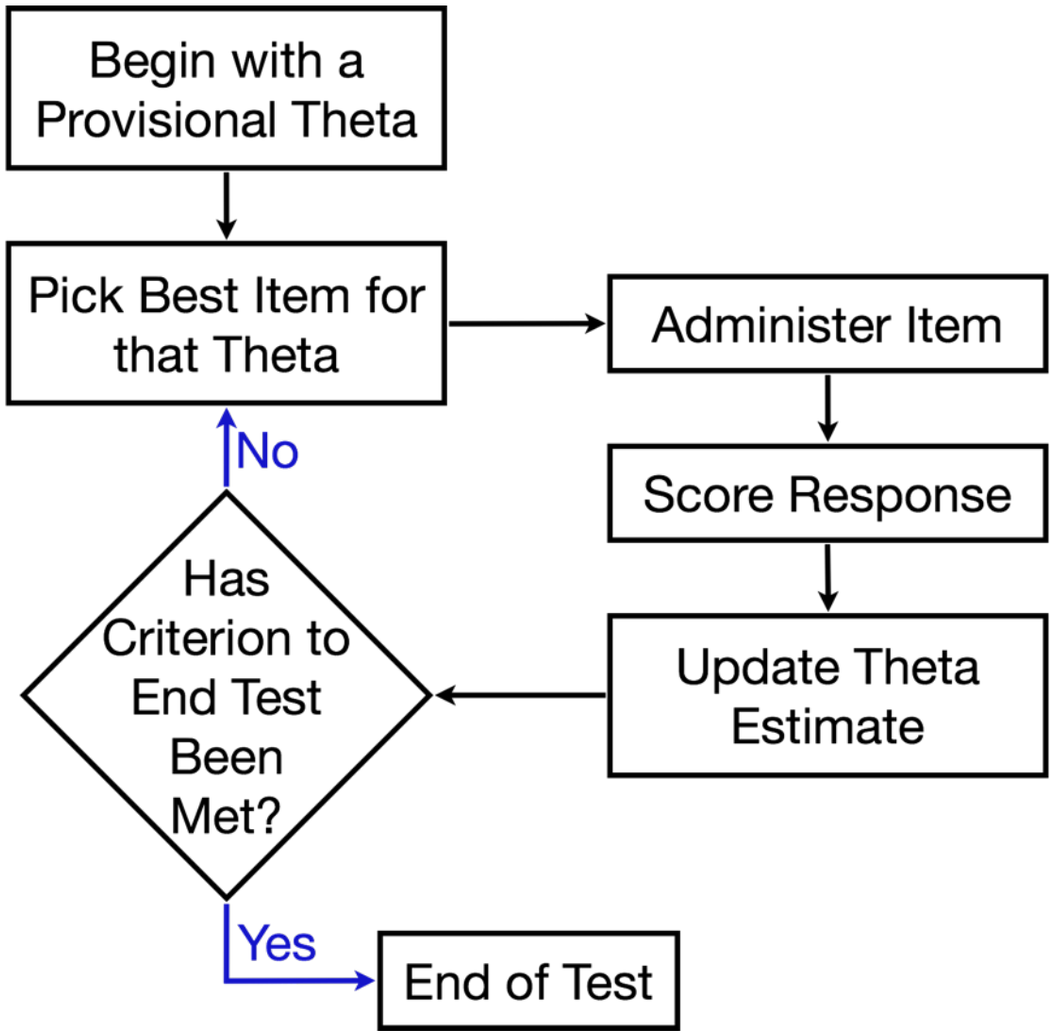
October is as close as possible in difficulty to the test they give in November. (You might remember this discussion when we talked about parallel tests – not that they could actually make it parallel, but they can do a pretty decent job). Conversely, with CAT we are actually trying hard to give two different people very different versions of the test. We want the high ability person to get a bunch of hard items and we want the low ability person to get a bunch of easy items. We actually want this. We want to violate this most revered rule of standardization. And yet, we'll still compare their scores head to head. Doesn't seem like it will work.

The answer is item response theory (IRT). With IRT, we can give different people totally different versions of the test (one full of the hardest questions and the other full of the easiest questions) and still compare the scores directly. We can do this because of the fact that we know the item parameters (difficulty, discrimination, and lower-asymptote) for all items and we score the test using maximum likelihood scoring. See Chapter 11 for a refresher on these issues. This of course means that we'll have to analyze the items to compute the item parameters (*b*, *a*, and *c*) for each item before we can set up a CAT.

The CAT Process

So, let's put those two pieces together: computers plus IRT. Now we have fully adaptive testing. We start the test knowing nothing about the test taker. After he or she answers the first item, we know a little something, but not much. Based on what we know, we can pick an item that might actually match his or her ability level. Once he or she answers the second item, we know a little more. The next item we pick should be an even better match to his or her ability level. And so on. We'll stop the test when we conclude that precision of our estimate of the test taker's ability will not be improved by the addition of more items. A flow chart of this process is presented in Figure 1.

FIGURE 1 The CAT Process



One issue we haven't discussed is the first step. It says we need to begin with a provisional theta estimate. We need something for theta in order to pick the first item (which would be the best item for that theta). We have a conundrum: We don't know what theta to use given that our test taker has not answered any questions yet. What to do. Hmmm… Let's just make up a theta and forget about it as soon as the test taker answers the first item. At that point we'll estimate theta for real. So we're going to have to just pull a theta out of thin air to start the test. We really have three options here, and two of them have problems. First, we could be optimists and say that everyone is high ability until they prove otherwise. That is, theta is +3.0 (or higher) until they actually miss an item. Sounds great, but problems abound. If I think your theta is 3.0, what kind of an item will I pick for you? That's right, a really, really hard one. It's not too fun to take a test and start with the hardest question. For some people, that can shake their confidence in ways that affect their performance on the rest of the test. Fine, so we'll go the other way, start people off with a low theta, like -3.0. No harm there since, we'll estimate theta for real after they answer the first item. This is nice

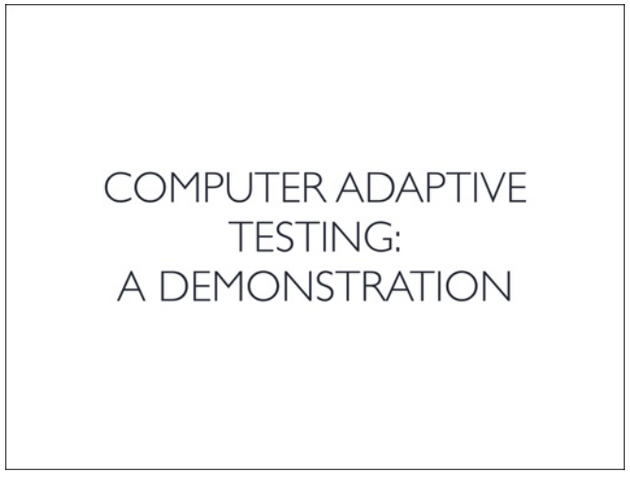
because the first item the test taker sees is very easy and may help build confidence. The only problem is that very few people actually need to take the easiest possible item – very few people are very low in ability. Thus, this item is a wasted item for most test takers. The solution must lie in the middle. We'll assume that everyone is average until they answer the first item. This way, the test starts with an item that is average in difficulty. Because most people are close to average, the first item will not be a wasted item for most people.

A Demonstration

Shown in Interactive 1 is an inside look at what happens when a computer adaptive test is administered. This test is a math test with a pool of 28 items. This test, like many tests, works best (is most precise) for thetas close to zero. We will observe the responses of an imaginary test taker. I'll report his theta and standard error for that theta after each item.

One thing that is clear is that our computer adaptive test could be improved if we were a little more judicious about our theta estimates early on in the test. Specifically, after one item the test taker's theta will be estimated to be either positive infinity (if the item was answered correctly) or negative infinity (if the item was answered incorrectly). These estimates cause the next item to be either the easiest or hardest item in our pool. This is obviously a ridiculous situation and not at all consistent with the goal of computer adaptive testing.

INTERACTIVE 1 Computer Adaptive Test Demonstration



The solution should be obvious. Rather than re-estimate theta after the first item has been answered, we should delay re-estimating theta until a small set of items (five, for example) have been answered by the test taker. Thus, the first real estimate of theta will be based on a larger sample of behavior and will be less likely to fluctuate so greatly. Essentially, our CAT doesn't start being a CAT until the test taker has has given us enough information for us to form a reasonable estimate regarding his or her ability.

Fifty Ways to Stop a CAT

The only remaining issue is: When do we stop the test? We have a lot of options there. What if we didn't stop the test? That is, we just keep giving items until we run out of items. Everyone takes all of the items. (Question: What do you call a computer adaptive test in which the test takers answer all of the items? Answer: a test.) In such a situation, we don't have an adaptive test. Sure, the order of the items is different for different people, but at the end of the day, everyone answers all of the items. Including the ones that are a waste of time (eventually, the high ability test taker must answer the easiest items in the shallow end of the item pool). If we wanted to give all of the items to everyone, we could have saved ourselves a lot of trouble and just given a standard, paper-and-pencil test. My point is, the stopping rule is one of the things that makes a CAT administration better than a traditional testing format. We don't want to stop too early (too few items) or too late (too many items). We've already talked about problems with giving too many items. But what if we give too few items? If we give too few items, our estimate of the test taker's theta is not as precise as it should be. Specifically, the standard error for this person's theta estimate will be too high. We'll explore some common stopping rules and see that most stopping rules result in too few items being

given to some test takers and too many to other test takers. Not very efficient.

The simplest stopping rule is to stop the test after a set number of items. If we decide that the magic number of items is 20, then everyone who takes the test will be done after they've answered the 20th question. High ability test takers will, in large part, have answered the 20 hardest items, whereas low ability test takers will have seen the 20 easiest items. You get the idea. Easy rule to understand and implement. Here's the problem: Recall that most tests work best (i.e., are most precise) for average theta values. That is, they have many items that yield information at around theta of zero. But for theta values at the extremes (e.g., beyond +/-1.5), most tests have very few items that yield information. Now based on that fact, a simple "stop after X items" stopping rule will result in people with thetas at the extremes taking too many items and people at the middle area taking too few items. Not good. Sure, it's better than a non-adaptive test, but it's not all it can be. A more sophisticated version of this same rule is to stop when a person's standard error of theta reaches a fixed threshold (e.g., stop when SE of theta is below .5). Ultimately, the same problems will occur because, as we said before, most tests work best for people with average ability levels. The good news is that there are more sophisticated stopping rules that do a great job of balancing the competing demands of measurement precision and test administration time.

CAT and Test Security

CAT can increase test security as compared with paper-and-pencil or computer based tests (CBT, a non-adaptive test which is administered on a computer). With a standard test, a test taker could memorize the items and give them to another person who will be taking the same form of the test in the future. It may sound far-fetched, but it's been done (ETS v. Kaplan, 1997).

Now let's see if this same security breach could happen with the CAT GRE. Let's say for the sake of argument that the CAT GRE item pool contains 300 hundred items for each section of the test. (Although I have close to zero inside information, it is my suspicion that the CAT GRE item pool is much greater than 300. It is likely to be well into the thousands.) Let's say that twenty bogus test takers take the test with each person having the goal of memorizing five items from each section of the test. That's 100 items from each section – a third of the test. Now let's say these items are shared with a new test taker. When he takes the test, he'll know the answer to about one of every three items. If we were scoring this test with number-right scoring, that would be a big boost. But this is CAT with IRT based maximum likelihood scoring. I am sure that you remember from Chapter 11 that with maximum likelihood scoring, an aberrant response (missing easy items but getting a hard item correct) is ignored to a certain extent. Now think about our cheating test taker who knows one-third of the items. Unless he is of high ability, he will have a seriously aberrant response pattern in which he will get some hard items right (because he was told about the item in advance) and miss a number of medium difficulty items that we did not see in advance.

But there's more. As long as we're using a computer to store our items, why not store a ton of items? As in a couple of thousand per section. Twenty test takers memorizing five items only gets you a total of one hundred – only five percent of the test. Not much help for Mr. Cheating Test Taker*, particularly when maximum likelihood scoring is involved. And there's more yet. What if we organized our 2000 items per section into clusters of 20 items per cluster (all items within a cluster sharing the same characteristics) and set a rule that a given test taker can only see one of the 20 items in a given cluster (chosen at random)? This means that even if our cheating test taker had a

perfect ability match to the test takers who stole the items, he would never see more than 5% of them. All of this is a long way of saying that with CAT, maximum likelihood scoring, and very large item pools, we can achieve much greater test security than was possible with other forms of test administration.

*Actually his real name

Final Thoughts

It occurs to me that we haven't discussed why anyone would want to use a CAT. Obviously, it's cool to pick just the right items for our test takers. So that's something you can to brag to your friends about. But there have to be better reasons than just that. What do we get out of tailoring test content to the ability of the test taker? There's also the test security angle. That's moderately cool. Anything else? Let's put it all together. If we have fifty items, and we only administer the useful ones (say, the 20 items that are well matched to the test taker's ability), then our test takers have taken a shorter test. Shorter in terms of items and time. And by not presenting useless items, we have reduced the opportunities for our test takers to make random errors (a low ability test taker can't guess correctly on a difficult item if we don't present him with that difficult item). So we can actually reduce the number of errors made by our test takers. Now that's seriously cool.

References

13

Because I didn't just make up all of this stuff.

References

AERA, APA, & NCME, (1999). *The Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Anastasi, A. (1982). *Psychological Testing (5th ed.)*. New York, NY: MacMillan Publishing Company.

Brown, R. D., & Cromwell, B. (2005). *Average intertest intervals used in test-retest reliability studies of published tests.* Unpublished manuscript.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich: Philadelphia.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335-337.

Ebel, R. L. (1975). *Prediction? Validation? Construct validity?* Paper presented at Content Validity II conference. Bowling Green, OH.

ETS v. Stanley Kaplan, 965 F. Supp 731 (D. Md. 1997).

Gasperson, S. M., Bowler, M. C., Wuensch, K. L., & Bowler, J. L. (2013). A statistical correction to 20 years of banding. *International Journal of Selection and Assessment, 21,* 46-56.

Guion, R. M. (1977). Content validity - the source of my discontent, *Applied Psychological Measurement, 1,* 1-10.

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77,*161-167.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society.

Lord, F. M., & Novick (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw Hill.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Earlbaum.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Stamp, J. (1929). *Some economic factors in modern life*. London: PS King & Son.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677-680.