CHAPTER 4

Conceptualization and Measurement

Theoretical Constructs
- Methods of Measurement
- Operationalizing and Measuring Constructs
- Advantages of Multiple Methods

Levels of Measurement
- Nominal Level of Measurement
- The Special Case of Dichotomies
- Ordinal Level of Measurement
- Interval Level of Measurement
- Ratio Level of Measurement
- Comparison of Levels of Measurement
- Why Measurement Level Matters

The Special Case of Psychometrics
- Evaluating Measures
- Measurement Reliability
- Stability Over Time

Internal Consistency
- Increasing Reliability
- The Logic of Measurement Validity
- Face Validity
- Content Validity
- Criterion Validity
- Construct Validity
- Latent Variable Models
- Culture and Measurement

Conclusions
- Key Terms
- Research in the News
- Activity Questions
- Review Questions
- Ethics in Action

©2015 SAGE Publications
LEARNING OBJECTIVES: FAST FACTS

Measurement in the Testing of Ideas

- A construct is simply an idea, concept, or theoretical abstraction that has multiple referents, which in psychology are known as measurable variables.
- To “play” with theoretical constructs such as intelligence, happiness, and personality requires objective instruments of measurement. No one measurement can fully capture a construct.
- A tool of measurement is judged by the extent to which it provides reliable (consistent) and valid (meaningful) scores.
- The correlation coefficient statistic is used to assess reliability and validity.
- A reliable instrument is internally consistent and stable over time, as reflected by high interitem and test–retest reliability coefficients.
- An instrument has convergent validity when it correlates with other measures of the same construct, and it has discriminant validity when it does not correlate with the measures of a different construct.

TESTING BEFORE LEARNING

Try your hand at these questions. Don’t worry about not knowing the correct response, as you haven’t read the chapter yet! But research shows that a pretest such as this can enhance learning (e.g., Kornell, Hays, & Bjork, 2009). So here is the answer. You come up with the question.

1. Best described as a theoretical abstraction
   a. What is a variable?
   b. What is a hypothesis?
   c. What is a sample?
   d. What is a construct?

2. No single variable can fully capture the breadth and depth of my meaning.
   a. What is a variable?
   b. What is a hypothesis?
   c. What is a sample?
   d. What is a construct?

3. Tells you how consistent your measure is, both internally and over time
   a. What is reliability?
   b. What is validity?
   c. What is bias?
   d. What is a construct?

4. Tells you how good your measure is, that is, the extent to which it captures the concept (e.g., intelligence) for which it was designed
   a. What is reliability?
   b. What is validity?
   c. What is bias?
   d. What is a construct?

5. Especially important for statistical analyses
   a. What is a construct?
   b. What is a variable?
   c. What is level of measurement?
   d. What is a hypothesis?

6. A good example of a ratio level of measurement because I have an absolute zero point with fixed measuring units
   a. What is income?
   b. What is intelligence?
   c. What is personality?
   d. What is anxiety?

7. A level of measurement that is used with categorical data, such as gender, team membership, geographical location, yes/no responses, etc.
   a. What is a nominal level of measurement?
   b. What is an ordinal level of measurement?
Imagine that you have been asked to measure something that is dear and close to all our hearts—our happiness! To whet your appetite, please read each of the following five statements, and indicate the degree to which you agree with each statement by recording a number between 1 (strongly disagree) and 7 (strongly agree).

1. In most ways, my life is close to my ideal.
2. The conditions of my life are excellent.
3. I am satisfied with my life.
4. So far, I have gotten the important things I want in life.
5. If I could live my life over, I would change almost nothing.

Now compute your total score by adding up your five responses. The psychologist Ed Diener and his colleagues have given this test, known as the Satisfaction With Life Scale (SWLS), to thousands of people (Diener, Emmons, Larsen, & Griffin, 1985; Pavot & Diener, 1993). Their results indicated that people tend to report being generally satisfied with their lives. In fact, as Pavot and Diener (1993) reported, most samples fall in the range on the SWLS of 23 to 28, meaning that the average response in most groups to the individual questions was about 5 or 6 on the 7-point scale, which Pavot and Diener interpreted as indicating slightly satisfied to satisfied.

The SWLS is used as a measurement tool by psychologists to investigate the age-old question “What is happiness?” Long the interest of philosophers and wise laypersons, the question of what is happiness has emerged as a fertile area of research in psychology. Indeed, the past 20 years have witnessed the emergence of a school of psychology known as positive psychology. Defined as the scientific study of ordinary human strengths and virtues, positive psychology seeks to learn about human thriving, flourishing, optimism, resilience, joy, and capabilities (Seligman, 1991; Sheldon & King, 2001). In short, positive psychology is interested in studying happiness.
Theoretical Constructs

In positive psychology, subjective well-being is the term used for what we refer to in everyday language as happiness (Diener, 2000). Subjective well-being is what is called a theoretical construct. A construct is an idea, concept, or abstraction that has multiple referents, which in psychology are known as measurable variables. A construct does not exist as a material entity; it is not concrete as, say, a table is concrete; it cannot be held or touched; it cannot be hit, kicked, or scratched.

A psychological construct such as “subjective well-being” can be measured by a variety of techniques or operations. For example, anxiety is a psychological construct that can be measured in several different ways. It can be measured by objective self-report tests that ask questions about specific behaviors, feelings, or thoughts that are theorized to be indicators of anxiety. It can be measured when a researcher uses observational techniques to code particular behaviors. It can also be measured physiologically by recording galvanic skin response, that is, changes in the electrical properties of the skin.

A construct is cast in a psychological theory. Let’s imagine that the SWLS you just completed piqued your interest. You now have become passionate about studying the question “What is happiness?” In your research, you discover the construct of subjective well-being. You learn about the psychological theory that describes, defines, and explains the construct of subjective well-being. The idea, perhaps to your surprise, is quite simple. You learn that subjective well-being is defined conceptually as the extent to which people view themselves as living good lives (Diener, 2000). In other words, each person is the final arbiter or judge of the quality of his or her life (Diener, 2000). With this subjective definition in mind, you can surely see why the SWLS is such a widely used tool in studies of happiness. Indeed, from your experience from taking the test, you no doubt know that the SWLS is all about having people evaluate their own lives using their own individual standards.

Note how theory provides a conceptual framework to study a particular topic. Let’s think of a different theoretical approach to the question of well-being. Perhaps you find the SWLS to be too subjective. Well-being, you theorize, goes beyond simple personal satisfaction that is measured with the SWLS. Your theoretical approach is more clinical than that of the subjective well-being studies pioneered by Diener and colleagues. Instead of subjective well-being, your research leads you to the construct of mental health. You discover the work of the social scientist Corey Keyes (2007), who offers a conceptual definition of mental health that you find intriguing: “a state of successful performance of mental function, resulting in productive activities, fulfilling relationships with people, and the ability to adapt to change and cope with adversity” (U.S. Public Health Service, 1999, p. 4).

Thus, here we have two approaches from the same school of psychology, each essentially researching the question “What is happiness?” For Diener and colleagues, their theoretical approach is organized around the psychological construct of subjective well-being. For Keyes, it is the construct of mental health for which he adopted a clinical perspective, and in fact developed a classification system of mental health that parallels the Diagnostic and Statistical Manual, used for mental illness.

Methods of Measurement

Variables are intimately connected to constructs. Think of a variable as derived from a construct; that is, SWLS is a variable derived from the construct of subjective well-being (Pavot & Diener, 1993). However, no one variable alone can fully capture a theoretical construct. In other words, no single tool of measurement can by itself fully capture a construct.
To underline this point, consider the experience sampling method (ESM), another tool that is used to assess subjective well-being (Csikszentmihalyi & Csikszentmihalyi, 1988; Csikszentmihalyi & Larson, 1987; Krueger & Schkade, 2008). The ESM provides a record of current circumstances and feelings as they are experienced in real time. Its main objective is to measure the subjective experiences of persons as they are engaged in their everyday lives. Let’s imagine that you are a participant in a study that uses ESM. Typically, you would be prompted or signaled by an electronic device such as a cell phone at irregular times within a defined period (e.g., between 9:00 a.m. and 9:00 p.m.). In response to the prompt, you would record what you are experiencing and feeling at that time.

For example, the sample of 2,250 adults in an ESM study by Killingsworth and Gilbert (2010) received a signal on their iPhone at random times; the signal asked them to respond to three simple questions: a happiness question (“How are you feeling right now?”); an activity question (“What are you doing right now?”); and a mind-wandering question (“Are you thinking about something other than what you are currently doing?”). Their results indicated that people spend considerable time mind wandering, which in turn caused them to feel unhappy. The researchers concluded that “a human mind is a wandering mind, and a wandering mind is an unhappy mind” (p. 932).

Notwithstanding the exciting prospects of the marriage of ESM with smart phone technology, there have also been developments in more traditional, low-tech, pencil-and-paper measures of subjective well-being. Enter the day reconstruction method (DRM). Here, you would be asked to reconstruct the experiences that you had yesterday (Kahneman et al., 2004). That is, participants are instructed to think about the preceding day, which is divided into episodes, and to describe each episode. Kahneman and colleagues wanted to know how the DRM stacked up against the gold standard of the EMS. Their results were very encouraging for the use of the DRM. In particular, they found that DRM reports from 909 employed women corresponded with the results previously established with research using the ESM.

**Operationalizing and Measuring Constructs**

Each of these three approaches, the SWLS, the ESM, and the DRM, provides a distinct operational definition for the construct of subjective well-being. As you know, the SWLS is quick, easy to administer, and easy to take. By contrast, the DRM takes between 45 and 75 minutes to complete, but it provides data beyond what is collected using the five-item SWLS, including how much time respondents spend doing certain activities and how they feel when they are engaged in these activities. In fact, the DRM is intended to reproduce the information that the ESM captures in real time. Its key advantage over the ESM is that instead of providing only a sampling of moments, DRM offers an assessment of contiguous episodes over a full day (Kahneman et al., 2004). Now let us see how to use scores from the SWLS and the DRM to operationalize subjective well-being.

First for the SWLS, take your score along with your classmates’ scores. This would give you a sample of scores. Scores can range as low as 5 and as high as 35. As you can see by the items, a lower score would indicate dissatisfaction with life, and a higher score would indicate satisfaction with life. In particular, a score of 20 represents the neutral point on the scale, meaning that the respondent is about equally satisfied and dissatisfied with life. Pavot and Diener (1993) provided a useful descriptive classification for SWLS scores: 26 to 30 = satisfied, 21 to 25 = slightly satisfied, 15 to 19 = slightly dissatisfied, and 5 to 9 = very dissatisfied with life.

And as we know from Chapter 2, sample scores can be summarized using descriptive statistics of central tendency (e.g., mean) and variation or range (e.g., standard deviation). In fact, as Pavot and Diener (1993) reported, most samples fall in the range on the SWLS of 23 to 28, meaning that most groups range from slightly satisfied to satisfied. For the SWLS, you could collect the scores and then tabulate the class average and standard deviation. You could compare your summary statistics with those reported by Pavot and Diener (1993). How does the level of life satisfaction for your class compare with the findings of Pavot...
and Diener (1993)? And yes, by the way, congratulations, you have successfully operationalized subjective well-being using the SWLS!

The DRM also provides a means to operationalize subjective well-being. It assesses “experienced well-being,” in contrast to the SWLS, which measures general life satisfaction. These two methods highlight two aspects of well-being. The DRM hopes to capture the well-being that people experience as they live their lives (Kahneman, 2011). The other aspect, for which the SWLS is designed, relates to the issue of how people evaluate their lives, that is, the judgment that they make when asked to evaluate their lives (Kahneman, 2011). The important point is that experienced well-being and life evaluation each represent two components of the construct of subjective well-being. That each is assessed and operationalized using different methods of measurement underscores the critical point that no one variable can fully capture the meaning of a construct.

Kahneman (2011) recently reflected on the roles of experienced well-being, as measured by DRM, and life evaluation, as indexed by SWLS. He noted that life evaluation and experienced well-being can be influenced by different events. For example, educational achievement is associated with greater life satisfaction but not with greater experienced well-being. In fact, Kahneman (2011) noted that in the United States, educational achievement may be linked to higher reported levels of stress. In a similar vein, adverse health detracts more from experienced well-being than from life evaluation. As Kahneman (2011) concluded, experienced well-being and life evaluation are related but different aspects of happiness. And we will give Kahneman the last word on what is the good life, as he put it best: “It is only a slight exaggeration to say that happiness is the experience of spending time with people you love and who love you” (p. 395).

Advantages of Multiple Methods

Overall, keep in mind that the variables and particular measurement operations chosen for a study should be consistent with the research question. For example, Keyes chose variables and particular measurement operations that were aimed toward investigating and addressing the question of whether mental health reflects more than simply the absence of mental illness. To address these questions, he used two kinds of methods of measurement: objective self-report adjective checklists and the interview. These two methods of measurement provided him with data to demonstrate that mental health means more than the absence of mental illness. As an example, adults with mental illness had higher levels of positive functioning with regard to life goals, resilience, and friendships than adults without mental illness but who were classified as “pure languishing.” On the basis of these findings, Keyes (2007) concluded that languishing could often be worse than the presence of a diagnosable mental illness.

Multiple methods of measurement are thus essential for us to develop a complete understanding of a phenomenon, such as the relationship between mental health and mental illness, or for distinguishing different aspects of happiness, such as life satisfaction and experienced well-being. Keep in mind the importance of using different forms of measurement of the same construct. By doing so, you can test whether these multiple methods of measurement are related, and the degree to which they are adds to our understanding of a given construct. We will discuss more on this later in the chapter.

Levels of Measurement

Measurement is the foundation of psychological research. It is defined as the process by which numbers are used to designate objects or events according to some objective rule. By rule, we mean, the way numbers are used to represent a particular scale or level of measurement. A variable can have one of four scales
of measurement: (1) nominal, (2) ordinal, (3) interval, or (4) ratio (see Exhibit 4.1). The operational
definition of a variable specifies the particular level of its measurement. Knowing the level of measurement
of a variable dictates the mathematical and statistical procedures that can be computed. As such, levels of
measurement are important to understand, particularly for data analysis.

<table>
<thead>
<tr>
<th>Exhibit 4.1 Four Levels of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
</tr>
<tr>
<td>Ordinal</td>
</tr>
<tr>
<td>Interval</td>
</tr>
<tr>
<td>Ratio</td>
</tr>
</tbody>
</table>

**Nominal Level of Measurement**

The **nominal level of measurement** (also called the categorical or qualitative level) identifies variables whose values have no mathematical interpretation; they vary in kind or quality but not in amount. In fact, it is conventional to refer to the values of nominal variables as “attributes” instead of values. “State” (referring to the United States) is one example. The variable has 50 attributes (or categories or qualities). We might indicate specific states with numbers, so California might be represented by the value 1 and Oregon with the value 2 and so on, but these numbers do not tell us anything about the difference between the states except that they are different. Oregon is not one unit more of “state” than California, nor is it twice as much “state.” Nationality, occupation, religious affiliation, and region of the country are also variables measured at the nominal level. Put simply, a categorical variable is one that uses a nominal scale of measurement (see Exhibit 4.2).

Although the attributes of categorical variables do not have a mathematical meaning, they must be assigned to cases with great care. The attributes we use to measure, or categorize, cases must be mutually exclusive and exhaustive:

- A variable’s attributes or values are **mutually exclusive** if every case can have only one attribute.
- A variable’s attributes or values are **exhaustive** when every case can be classified into one of the categories.

When a variable’s attributes are mutually exclusive and exhaustive, every case corresponds to one, and only one, attribute. We know this sounds pretty straightforward, and in many cases it is. However, what we think of as mutually exclusive and exhaustive categories may really be so only because of social convention; when these conventions change, or if they differ between the societies in a multicountry study, appropriate classification at the nominal level can become much more complicated.

**The Special Case of Dichotomies**

**Dichotomies**, variables having only two values, are a special case from the standpoint of levels of measurement. The values or attributes of a variable such as gender clearly vary in kind or quality but not in amount.
Thus, the variable is categorical—measured at the nominal level. Yet we can also think of the variable as indicating the presence of the attribute “female” (or “male”) or not. Viewed in this way, there is an inherent order: A female has more of the “female” attribute (it is present) than a male (the attribute is not present). So what do you answer to the test question “What is the level of measurement of ‘gender’?” “Nominal,” of course, but you’ll find that when a statistical procedure requires that variables be quantitative, a dichotomy can be perfectly acceptable. As we can see from the graphs (see Exhibit 4.3), changing the number assigned to “male” and “female” does not change the number of men and women participants in a study. We still have the same number of men and women in the data set.

In Chapter 2, we created a dichotomous variable when we categorized participants as either maximizers or satisficers. As you recall, we categorized respondents based on their scores on the measure of maximization. Note that the scores on the maximization measure represent quantitative data that we used to create these two qualitative categories. Scores that fell below the cutoff on the maximization measure were categorized as satisficers and those above the cutoff on the maximization measure were categorized as maximizers.

Now let’s imagine that we want to compare maximizers and satisficers on grade point average. We would create a dichotomous variable that we call “decision-making style,” which would have two attributes, satisficer (made up of all respondents who scored below the cutoff) and maximizer (made up of all respondents who scored about the cutoff). We could code “1” for satisficer and “2” for maximizer. These numbers have no arithmetical meaning, but the numerical code needs to be consistent, for this example, all satisficers must be coded with “1” and all maximizers must be coded with “2.” Once we have our dichotomous code in place to categorize maximizers and satisficers, we could then use software-based statistical analyses to compare the two groups on grade point average.

**Ordinal Level of Measurement**

The first of the three quantitative levels is the **ordinal level of measurement**. At this level, the numbers assigned to response choices allow for “greater than” and “less than” distinctions. And because this level of
measurement allows for a ranking of responses, such as degree of agreement or frequency of occurrence, ordinal level of measurement is sometimes referred as a ranking scale.

Oftentimes, research participants are asked to rank the frequency of occurrence of a particular emotion or behavior on a numerical scale. For example, Keyes asked participants to rate how frequently in the past 30 days they felt (a) cheerful, (b) in good spirits, (c) extremely happy, (d) calm and peaceful, (e) satisfied, and (f) full of life. For each of these positive emotions, participants rated how frequently they had experienced each on a scale by selecting only one of the five numbered responses (typically presented from left to right) with 1 = none of the time, 2 = a little of the time, 3 = some of the time, 4 = most of the time, or 5 = all of the time.

Researchers commonly referred to this metric as a Likert scale, named after its originator, psychologist Rensis Likert (Likert, 1932). Keyes used another ordinal scale when he asked participants to “rate their life overall these days” on a scale ranging from 0 = worst possible life overall to 10 = best possible life overall.

A simple depiction of an ordinal scale of measurement is presented in Exhibit 4.4. Here imagine that the respondents participated in a driving test in which they rank ordered four different cars on a 1 to 4 scale, with 1 = most favorite and 4 = least favorite. Exhibit 4.5 simply shows that ranking always involves an ordinal scale, such as that used by Olympic judges. Other common examples of an ordinal level of measurement are letter grades, military rank, and socioeconomic status.

As with nominal variables, the different values of a variable measured at the ordinal level must be mutually exclusive and exhaustive. They must cover the range of observed values and allow each case to be assigned no more than one value. Often, instruments that use an ordinal level of measurement simply ask respondents to rate their responses to some questions or statements along a continuum of, for example, strength of agreement, level of importance, or relative frequency.

A key limitation of the ordinal level of measurement is that you cannot assume that respondents perceive the differences between response scale points as equidistant. For example, you cannot assume that the distance or difference between “all of the time” and “most of the time” is equal to the distance or difference between “a little of the time” and “none of the time.” While these points on the response scale are intended to
reflect an order of frequency ("all of the time" to "none of the time"), the numbers that you would assign to each of the response scale points (e.g., all of the time = 5, most of the time = 4, some of the time = 3, a little of the time = 2, and none of the time = 1) do not indicate the magnitude of difference between any two points on this ordinal response scale.

For instance, say you answered "all of the time," which is assigned and coded with a value of 5 for how frequently you felt cheerful in the past 30 days, and another participant responded with "none of the time," which we coded with a value of 1. You cannot say that the "all of the time" response is five times greater than the "none of the time" response just because you decided to code these responses as 5 and 1, respectively. The SWLS that you took at the beginning of this chapter uses an ordinal scale of measurement.

**Interval Level of Measurement**

An *interval level of measurement* has all the characteristics of nominal and ordinal scales of measurement. That is, like a nominal scale, it gives a name or category for each observation, with the number serving as a code for a label (e.g., 1 = females, 2 = males). As in an ordinal scale, responses are numerically ordered or ranked from lowest to highest on some particular characteristic. But in addition, an interval level of measurement by definition uses a scale on which the distances between any two points are of known size. The numbers indicating the values of a variable at the interval level of measurement represent fixed measurement units but have no absolute, or fixed, zero point.

What this means is that a zero value on an interval scale does not indicate the complete absence of the measured variable. This is true even if the scaled values happen to carry the name "zero." The Fahrenheit scale illustrates this issue (see Exhibit 4.6). A temperature of 0°F Fahrenheit does not represent the complete absence of temperature, defined as the absence of any molecular kinetic energy. Rather, 0°F Fahrenheit became a handy convention for temperature measurement for largely accidental reasons of history.

The key point is that because an interval scale has no true zero point, ratio measurements make no sense. For example, we cannot state that the ratio of 40°F to 20°F Fahrenheit is the same as the ratio of 100°F to 50°F. After all, if the “zero” degree applied at the temperature that Fahrenheit happens to measure as 10°F, the two ratios would instead be 30 to 10 and 90 to 40, respectively, and would no longer be equivalent. For this reason, that is, the absence of a true zero point in an interval scale, we cannot say that 80°F is “twice as hot” as 40°F. Such a claim would assume a true zero point, which an interval scale lacks. Remember, as with all interval scales, that decision about where to “start” the Fahrenheit temperature scale is arbitrary and is not tied to an underlying physical reality of atmospheric climate.

In psychology, standardized measures of intelligence, personality, and aptitude typically use an interval level of measurement, just as do education tests such as the SAT (Scholastic Assessment Test) and GRE (Graduate Record Examination). That is, these standardized measures use an arbitrary zero point, as there is, for example, no such thing as no intelligence, no personality, and the like. Likewise, as with our example of temperature, you cannot say an IQ (Intelligence Quotient) score is twice as large as another IQ score. So while an IQ of 130 is 65 scaled points greater than an IQ of 65, it is not twice as great.

**Ratio Level of Measurement**

The numbers indicating the values of a variable at the *ratio level of measurement* represent fixed measuring units and an absolute zero point (zero means absolutely no amount of whatever the variable indicates).
Income is measured on a ratio scale, as no income equals zero dollars and cents! Weight is also measured on a ratio scale, as the scale of ounces and pounds has a genuine zero point. Kelvin temperature is another example of a ratio scale because in classical physics, zero on the Kelvin scale means absolute zero, the case in which all motion stops. Speed in terms of time in track and field races would also be considered to be measured on a ratio scale (see Exhibit 4.7).

For most statistical analyses in social science research, the interval and ratio levels of measurement can be treated as equivalent, but there is an important difference: On a ratio scale, 10 is 2 points higher than 8 and is also two times greater than 5. Ratio numbers can be added and subtracted, and because the numbers begin at an absolute zero point, they can be multiplied and divided (so ratios can be formed between the numbers). For example, people’s ages can be represented by values ranging from 0 years (or some fraction of a year) to 120 or more. A person who is 30 years old is 15 years older than someone who is 15 years old (30 - 15 = 15) and is twice as old as that person (30/15 = 2). Of course, the numbers also must be mutually exclusive and exhaustive so that each case can be assigned one and only one value.

### Comparison of Levels of Measurement

Exhibit 4.8 summarizes the types of comparisons that can be made with different levels of measurement as well as the mathematical operations that are legitimate. All four levels of measurement allow researchers to assign different values to different cases. Scores obtained on ordinal, interval, or ratio scales of measurement all allow for mathematical operations and are thus often referred to as quantitative variables.

<table>
<thead>
<tr>
<th>Examples of Comparison Statements</th>
<th>Appropriate Math Operations</th>
<th>Relevant Level of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A is equal to (not equal to) B</td>
<td>= (≠)</td>
<td>XX XX XX XX</td>
</tr>
<tr>
<td>A is greater than (less than) B</td>
<td>&gt; (&lt;)</td>
<td>XX XX XX XX</td>
</tr>
<tr>
<td>A is three more than (less than) B</td>
<td>+ (−)</td>
<td>XX XX</td>
</tr>
<tr>
<td>A is twice (half) as large as B</td>
<td>× (/)</td>
<td>XX</td>
</tr>
</tbody>
</table>

By contrast, a nominal level of measurement involves classifying observations using an arbitrary number as a code for group membership (e.g., females = 1, males = 2). Thus, a nominal level of measurement does not allow for mathematical operations and is used for what are often referred to as qualitative variables. Keep in mind that researchers choose levels of measurement in the process of creating operational definitions for their variables; the level of measurement is not inherent in the variable itself. Many variables can be measured at different levels with different procedures.
Why Measurement Level Matters

Levels of measurement are especially important for statistical analyses that we will cover in depth in Chapter 10. We will learn in Chapter 10 that different types of statistical procedures require different levels of measurement of variables. For example, nonparametric statistics test hypotheses for variables that use either a nominal or an ordinal scale of measurement.

A commonly used nonparametric statistical procedure that will be covered in Chapter 10 is the chi-square test. Let’s imagine for a moment that you are interested in the question of whether there are more women than men majoring in psychology. To answer this question, you could perform a chi-square statistic test comparing percentages of female and male psychology majors. You would use the nonparametric chi-square test because gender is a categorical variable, meaning that it is measured on a nominal scale that can only be analyzed using nonparametric statistics.

On the other hand, parametric statistics are used with variables that are measured on either an interval or a ratio scale. A commonly used parametric statistical procedure that will be covered in Chapter 10 is the t-test statistic. Imagine you wanted to compare the SATs of female and male college applicants. A t-test would allow you to compare the average SATs of these two groups. The SAT would be considered a variable that is measured on an interval scale and thus suitable to be analyzed with parametric statistics, such as the t-test.

In general, parametric statistics are preferable because they provide a more powerful test than do nonparametric statistics of a research hypothesis. That is, parametric statistics provide the best chance of establishing whether there is a relationship between independent and dependent variables. Exhibit 4.9 depicts the relationship of levels of measurement and parametric and nonparametric statistics.

Exhibit 4.9 Levels of Measurement and Statistical Approaches

The Special Case of Psychometrics

Psychometrics is a school of psychology that studies the application of psychological tests as objective measures of the mind and mental processes. The word psychometric literally means “to measure the mind.” At the heart of the psychometric approach is that people have certain levels of psychological traits, such as those related to intelligence, personality, or attitudes, and that these presumed enduring and stable dispositions can be measured by objective tests. Equally important for the psychometric approach are competencies, such as those related to scholastic aptitude or academic achievement, which can also be measured by objective pencil-and-paper tests, such as the SAT (see Exhibit 4.10).

Psychometrics provides us with the most powerful set of tools and concepts for the measurement of human traits. These measurement tools or tests are used to operationalize constructs, such as...
psychological traits. In psychometrics, a specific set of procedures is used to operationalize a construct. These procedures are quite extensive and represent a field of study in and of itself in psychology and in education, known as test construction or tests and measurements (see Kline, 1986).

The psychometric approach to creating operational definitions uses the principle of standardization of both test administration and test scoring. Standardization simply means that testing, scoring, and interpretation procedures are uniform across all administrations of the test. In constructing a psychometric test, a researcher identifies a standardization sample, which is defined as a random selection of people drawn from a carefully defined population. As we will study in Chapter 5, random sampling simply means that each person of the targeted or defined population has an equal chance of being selected to be part of the standardization sample.

The tests, administered uniformly to the standardization sample, are then scored according to the same specified and objective criteria for each examinee. From the standardization sample, the norms of the test are calculated. The norms of a test involve two measures. The first is the mean or average score for the standardization sample. The second measure is the standard deviation, which indicates the degree to which an individual score deviates from the sample mean (see Chapter 10 for details).

In psychology, psychometrics has long been the dominant approach to the study of the construct of intelligence. Let us consider a widely used standardized test of IQ known as the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV; Wechsler, 2008). Test construction begins with a conceptual definition of the construct under study. In this case, Wechsler (1944) provided the original conceptual definition for the construct of intelligence as the “capacity of the individual to act purposely, to think rationally, and to deal effectively with his environment” (p. 3). This conceptual definition represents one of the earliest formulations of intelligence, but one that has remain controversial in psychology and in society in terms of its measurement of cultural meaning (see, e.g., Brown, 1992; Lemann, 1999; Neisser et al., 1996; Nisbett et al., 2012).
The most recent revision in 2008 of the Wechsler measure of intelligence offers a textbook example of the psychometric approach to test construction. First, based on an ever-growing body of empirical research, the conceptual definition of intelligence has been further developed and refined and now includes four distinct factors: (1) verbal comprehension, (2) perceptual reasoning, (3) working memory, and (4) processing speed. Test questions and tasks have been revised. Second, and in keeping with the psychometric approach, norms were updated for the WAIS-IV. Creating norms for a test is a defining feature of the psychometric approach. Norms help to standardize a test, and they require frequent updating, so that a test does not become obsolete. For the WAIS, updating test norms requires a new standardization sample. Given the rapid changes in society, the WAIS is updated approximately every 10 years. The next edition (fifth edition) of the test is due in 2018.

In "norming" the WAIS-IV, the developers gave the IQ test to a standardization sample of 2,200 English-speaking people who were randomly selected from a carefully defined population. The norming of the test came from a stratified sampling of the population. A **stratified sample** is formed by randomly selecting from relatively uniform subpopulations called strata. Strata are defined along specified, objective features. As part of the standardization process, the test scores generated from a standardization sample are plotted or graphed. The graph shows the range of scores as well as the distribution of scores, which is simply a tally of the number of examinees who obtained a certain score.

As we had covered in Chapter 2, this kind of graph (histogram) provides a frequency distribution for the scores generated from a sample. For many human traits, from height to weight to intelligence and personality predispositions, scores are distributed along a **normal curve** that is represented in Exhibit 4.11 (Kosslyn & Rosenberg, 2001). When the distribution of scores approximates a normal curve, most scores fall near the center, with gradually fewer scores as you move toward either extreme. In general, the larger the sample, the more closely will the distribution of scores resemble the normal curve, which is also known as the **bell curve**, due its bell shape—bilaterally symmetrical with a single peak in the middle.

Exhibit 4.11 shows the normal curve for WAIS-IV IQ scores. WAIS-IV raw scores are converted by a specific statistical formula to standard scores, with a mean of 100 and a standard deviation of 15. As you can see in Exhibit 4.11, about two thirds of the scores from the standardization sample fall within 1...
standard deviation above and below the mean of 100. In other words, two thirds of all scores fall within the range of 85 to 115. What this means is that a person with an IQ of 115 earns a score that falls 1 standard deviation unit above the mean of 100. In qualitative terms, an IQ of 115 would be classified as falling in the high average range. Thus, you can see how a person’s score on a standardized test can be compared and interpreted relative to others in the standardized sample.

For a test to be meaningful or useful, it must be both reliable and valid. As we learned in Chapter 1, reliability refers to consistency, as for example, the extent to which a research finding can be replicated. In measurement, reliability reflects the consistency of a test. There are two types of reliability of a measure. One is test–retest reliability, which is used to assess the stability of scores over time on a measurement. Internal consistency of a measure is another type of reliability, which can be used when you want to assess, for example, the extent to which the items of a test cohere. This type of reliability can be very useful because high internal consistency could mean that the test is measuring a unitary trait. Validity, on the other hand, is different from reliability and refers to the meaning of what is measured. Validity can be evident by what a measure might predict, for example, IQ scores predict school success.

Multiple methods of measurement in the form of reliable and valid tests are essential for understanding a theoretical construct. Consider the construct of well-being. Recall the two measures, SWLS and DRM, used to assess well-being. Note that these two different measurements revealed important insights about well-being. To wit, findings from studies that have employed these measurements indicated that life satisfaction and experienced well-being may represent two distinct aspects of happiness. In large part, we owe this advancement in our understanding of happiness to the development and use of these two different measures that researchers have employed to assess the construct of well-being. This represents an example of good research practice, namely, the use of multiple, reliable, and valid measures. Now let us learn how we too can practice good research using reliable and valid measures.

**Measurement Reliability**

A reliable test yields consistent scores when the construct being measured is not changing (or the measured scores change in direct correspondence to actual changes in the construct). If a test is reliable, it is affected less by measurement error or chance variation than if it is unreliable. Reliability is a prerequisite for measurement validity: We cannot really measure a construct if the tests we are using give inconsistent results. In fact, because it usually is easier to assess test reliability than test validity, you are more likely to see an evaluation of measurement reliability in a research report than an evaluation of measurement validity.

**Stability Over Time**

In the “Stat Corner,” we show how the correlational statistic provides a clear and direct assessment of the reliability of a measure. This statistic is used to compute a reliability coefficient of a measure. The stability of a measure over time is assessed by what is referred to as the test–retest reliability coefficient. The test–retest reliability coefficient provides a quantitative index of the precision of the measurement, with a higher
correlation coefficient indicative of smaller measurement error. For example, a construct such as intelligence is presumed to remain stable over time. In fact, the actual test–retest reliability for the WAIS-IV full-scale IQ is a remarkable .96 (Wechsler, 2008).

A major shortcoming of test–retest reliability is the “practice effect,” whereby respondents’ familiarity with the questions from the first administration of the test affects their responses to the subsequent administration of the same test. As a result, many standardized tests have alternate forms, which come in very handy when an examinee might need to be retested over a relatively short period of time.

Think of a study that is investigating the effects of a new drug on memory. Participants are tested before and after a 60-day treatment trial. A control group that receives a placebo is also tested on two occasions, before and after receiving the inert substance. To control for practice effects that could influence test scores, alternate forms of the memory measurement would need to be administered. In selecting your memory measure, then, you would look at the alternate-forms reliability of the test. Alternate-forms reliability reflects the degree of agreement between two slightly different tests comprising different sets of questions or items of the same construct.

In a similar vein, often research involves different raters coding the same behavior. Here, consistency among raters needs to be established so that the same behavior is coded reliably. This type of reliability is referred to as interobserver reliability, which indexes the degree to which raters consistently code the same observed behavior. That is, interobserver reliability measures the degree of agreement among raters.

With two raters, the same correlation statistic used with test–retest reliability is acceptable for assessing interobserver reliability. However, with more than two raters, the intraclass correlation coefficient (ICC) is used to assess interobserver reliability. The statistical formula for the ICC is different from that for the correlation used for test–retest reliability. Yet the ICC is equivalent to what would be the average of the test–retest correlations between all pairs of tests (Shrout & Fleiss, 1979).

Internal Consistency

Tests usually consist of many items, all of which are typically intended to measure the same construct, whether IQ, a personality trait, or reading comprehension. When researchers use multiple items to measure a single construct, they must be concerned with interitem reliability (or interitem or internal consistency). The idea is that if the items of a test are all measuring the same construct, the correlation among these items should be high. The goal is to have consistency across items within a test.

One frequently used measure of internal consistency is the split-half reliability coefficient. Here, you randomly divide all items that are supposed to measure the same trait or construct into two sets. Calculate the total score for each randomly divided half. The correlation between the two total scales is the split-half reliability coefficient. Cronbach’s alpha is another reliability measure commonly used to assess interitem reliability. Very similar to split-half reliability, Cronbach’s alpha is computed by several iterations of randomly dividing a scale of items into sets of split halves. You (or the computer program) compute the reliability coefficient, recomputing on the next set, and so on until you have computed all possible split-half estimates of reliability.

Increasing Reliability

One rather simple and straightforward way to increase the internal consistency of an instrument is to make it longer! In other words, by adding more related items to your test, you will increase the internal consistency of the test. Why is this so?
In any response, there is random error of measurement—variation due to multiple little factors that have nothing to do with the construct you are trying to measure. Perhaps an answer is influenced by a story in that morning’s newspaper, by the noise of a truck passing by, and on and on. Any given response to a question will therefore contain a large component of random error of measurement. But as you increase the number of items in a test, the random error involved in each answer to each particular question will be cancelled out by the random error associated with the responses to the other questions. The more you add questions related to the concept of interest, the more the random error will wash out in this way.

Exhibit 4.12 compares test–retest reliabilities for three indexes commonly used to assess subjective well-being: (1) a single-item measure of general satisfaction, (2) the five-item SWLS, and (3) ESM-generated real-time reports of emotional state (Krueger & Schkade, 2008).

As you can see, the differences in test–retest reliabilities between the single-item measure and the five-item SWLS are striking. The single-item SWLS showed moderate reliabilities, falling between .40 and .66 (Krueger & Schkade, 2008). However, the longer, five-item SWLS had reliabilities as high as .84. In addition, as reported by Krueger and Schkade (2008), Eid and Diener (2004) used a powerful statistical technique known as structural equation modeling and estimated the stability of the five-item SWLS to be very high, around .90.

Exhibit 4.12  Reliability of Subjective Well-Being Measures

<table>
<thead>
<tr>
<th>Coefficient Alpha</th>
<th>Test–Retest</th>
<th>Time Interval</th>
<th>Construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-item measures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andrews and Whithey (1976)</td>
<td></td>
<td>.40–.66</td>
<td>1 hour Life satisfaction</td>
</tr>
<tr>
<td>Kammann and Flett (1983)</td>
<td></td>
<td>.50–.55</td>
<td>Same day Overall happiness, satisfaction</td>
</tr>
<tr>
<td>Multiple-item measures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alfonso and Allison (1992)</td>
<td>.89</td>
<td>.83</td>
<td>2 weeks SWLS</td>
</tr>
<tr>
<td>Pavot, Diener, Colvin, and Sandvik (1991)</td>
<td>.85</td>
<td>.84</td>
<td>1 month SWLS</td>
</tr>
<tr>
<td>Blais, Vallerand, Pelletier, and Briere (1989)</td>
<td>.79–.84</td>
<td>.64</td>
<td>2 months SWLS</td>
</tr>
<tr>
<td>Diener et al. (1985)</td>
<td>.87</td>
<td>.82</td>
<td>2 months SWLS</td>
</tr>
<tr>
<td>Yardley and Rice (1991)</td>
<td>.80</td>
<td>.50</td>
<td>10 weeks SWLS</td>
</tr>
<tr>
<td>Magnus, Diener, Fujita, and Pavot (1993)</td>
<td>.87</td>
<td>.54</td>
<td>4 years SWLS</td>
</tr>
<tr>
<td>ESM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steptoe, Wardle, and Marmot (2005)</td>
<td></td>
<td>.65</td>
<td>Weekend–weekday Experienced happiness</td>
</tr>
</tbody>
</table>

Note: SWLS = Satisfaction With Life Scale; ESM = Experience Sampling Method.

The take-home point is straightforward. Increasing the number of items of a measure improves test–retest reliability. In addition, Exhibit 4.12 also presents estimates of internal consistency (coefficient alpha) for the SWLS. As Exhibit 4.12 shows, the SWLS had highly internal consistency reliability, with coefficient alpha values ranging from .79 to .89. ©2015 SAGE Publications
The correlation coefficient is one of the most useful statistical tests in psychological research. Computed by a specific formula, this statistic provides an index of how closely related two variables are. This statistic, known as the correlation coefficient (symbolized as \( r \)), can range from .00 to +1.00 and .00 to \(-1.00\). If there is no relationship between two variables, that is, the variation in one variable has nothing to do with the variation in the other variable, then the correlation value will be .00 or close to .00. The closer a correlation between two variables is to 1.00, either +1.00 or \(-1.00\), the stronger is the relationship between the two variables. The positive and negative signs of the correlation coefficient tell the direction of the relationship between the two variables. A positive correlation coefficient (a “plus” sign) means that there is a positive linear relationship—as scores on one variable increase, so do scores on the other variable. A negative correlation coefficient (a “minus” sign) means that there is an inverse linear relationship—as scores on one variable increase, scores on the other variable decrease.

Researchers often use scatter diagrams to graph correlations. A scatter diagram is a graph of a correlation. A correlation always involves two variables. Let’s call one variable \( x \) and the other \( y \). Let us say a participant had a score of 5 on one variable, labeled \( x \), and a score of 10 on another variable, labeled \( y \). Make a two-dimensional graph with a horizontal \( x \)-axis and a vertical \( y \)-axis. Now plot the participant’s scores of 5, 10 (5 to the right and 10 up). Each participant can be represented by a point on the graph, in this example, 5, 10, which represents scores on the \( x \) and \( y \) variables. So if there are 12 participants, there should be 12 points. The direction of the points on the graph will signify whether the correlation between the two variables across all participants is either positive or negative. For a positive correlation, the graph will depict points that fall on an imaginary line that tilts upward to the right. For a negative correlation, the graph will depict points that fall on an imaginary line that tilts downward to the right.

The correlation coefficient is also used to quantify reliability. For example, test-retest reliability is the statistical correlation computed between scores on the same measures across two time periods for the same respondents. The correlation value or coefficient between the two sets of scores is used as the quantitative measure of test—retest reliability. Ideally, if the scores remained identical across the two testing intervals for all respondents, then the correlation coefficient would be 1.00. This would indicate perfect agreement between the scores for respondents across the two testing intervals. In reality, however, the scores of respondents vary if the test is administered again. As such, correlation coefficients of .70 to .80 are considered satisfactory values for test—retest reliability. The test—retest reliability correlation also reflects the extent to which the rank order of respondents’ scores is consistent across the two testing periods. Rank order simply means where or in which position in the class a respondent’s score falls in relation to the other test takers. If the rank order of scores for Test 1 is reproduced for Test 2, then the test—retest reliability will be high. By contrast, if the rank order of scores for respondents is all mixed up across the two testing sessions, then the test—retest correlation coefficient will be low.

Exhibit 4.13 presents a hypothetical set of scores for the same test given 1 year apart for the same nine respondents. The correlation between the two sets of scores is .95, which indicates very high test—retest reliability. This indicates that the scores on this test remained extremely stable over the 1-year period between Time 1 and Time 2. The scatterplot plots the scores for Time 1 and Time 2 for each of the nine respondents. As you can see, the imaginary line tilts upward, indicative of the strong positive correlation of .95.
The Logic of Measurement Validity

How do we know whether a test measures what it says it measures? Does the WAIS-IV measure intelligence? Does the SWLS measure well-being? In psychological measurement, these are questions about the validity of an instrument or a test.

A valid instrument is one that taps the construct we intend to tap. Put simply, a valid measure does what it is supposed to do—the WAIS-IV measures intelligence, the SWLS measures subjective well-being, and so on. Note that questions of validity are not the same as questions of reliability: Reliability and validity are not synonymous. An instrument may be very reliable because it provides a highly consistent measure of a construct, but it may be invalid because it measures the wrong construct.

To illustrate the difference between validity and reliability, consider the following example. Imagine you gave a standard IQ test like the WAIS-IV in English to a group of Swahili-speaking high school students in Kenya. You may find high test–retest reliability for the WAIS-IV, but would you have a valid measure of those students’ intelligence? No, you would have a measure of their knowledge of English, but you would not have a valid measure of their intelligence. For these Swahili-speaking students of Kenya, the WAIS-IV in English is a measure of English proficiency rather than a measure of intelligence.
Giving an English-language IQ test to assess the intelligence of Swahili-speaking Kenyan students is, of course, an obvious example of measuring a construct other than the one intended. However, in most instances in psychological measurement, the question of validity is subtler yet always present. Indeed, to a certain extent, all measures in psychology share this problem with validity. This, as we have learned, is because operational definitions of a construct are imperfect. Operational definitions inevitably include elements that are not supposed to be included, while portions of the underlying construct that should be measured are excluded.

This is why we need multiple measures of the same construct. We also need multiple indicators or outcome measures to assess the quality of a construct, such as the extent to which it explains and predicts certain phenomena. In other words, we need to assess the validity of a construct. Ultimately, we evaluate the validity of a construct by determining whether its multiple indicators cohere into predictable and logical patterns of relationships. In so doing, we consider various kinds of validity. We start with the simple question of face validity of an instrument.

**Face Validity**

The question of face validity asks whether a test “looks like” it measures what it is supposed to measure (Anastasi, 1988). Face validity is essentially a matter of opinion and lacks a firm scientific or empirical basis. It adds little or nothing to the validation of a test. In fact, if anything, problems can arise when what a psychological test is supposed to measure looks too obvious. Such an instrument may be subjected to social desirability bias, in which examinees respond defensively to test items based on how they wish to be perceived rather than openly and genuinely.

Other test-taking strategies, such as examinees “faking bad” or “faking good,” may also threaten the validity of an instrument whose measurement objective appears to be too transparent. Face validity is similar to the legal term *prima facie*, a Latin expression meaning “on its first appearance,” or “at first sight.” However, in law, prima facie evidence is an essential element of jurisprudence, whereas in psychology the face validity of an instrument is inconsequential.

For subjective well-being research, social desirability bias has long been viewed as a major concern. After all, well-being is a highly desirable state, and people may be inclined to be overly favorable in their judgments of life satisfaction, either, unconsciously, in order to preserve a positive self-image or, consciously, in order to project a good impression on others, including the researcher administering the SWLS.

Social desirability has long been considered a serious threat to the validity of any self-report measure. It is often measured using the Marlowe-Crowne Social Desirability Scale, which consists of 33 true/false items (Crowne & Marlowe, 1960). Consider two scale items: “I like to gossip at times” and “I am always willing to admit when I make a mistake.” Endorsing “false” for liking gossip and choosing “true” for always admitting a mistake are scored as socially desirable. A higher score raises the question of social desirability bias in the self-report. Good research practice is to include a measure of social desirability bias whenever self-reports are the critical source of data or evidence, as is indeed the case for subjective well-being research.

**Content Validity**

A valid test needs to provide a full measure of a construct. For example, the WAIS-IV includes 10 core subtests, each of which is presumed to measure a distinct aspect of intelligence, such as verbal reasoning or constructional abilities of working with your hands. Thus, content validity refers to the extent to which test items have sufficient breadth to capture the full range of the construct intended to be measured.

In this regard, Keyes (2007) went to great lengths to make sure that his measures provided the full range of the meaning of the construct of mental health. Recall that he used multiple measures to capture what he conceptualized to be the multidimensional nature of mental health. In so doing, his measurement of mental health as well as his DSM measure of mental illness would be considered to be valid from the standpoint of content validity.
Criterion Validity

The validity of a test can also be measured against a certain criterion or outcome. For example, a measure of blood alcohol concentration or a urine test could serve as the criterion for validating a self-report measure of drinking, as long as the questions we ask about drinking refer to the same period of time. Thus, criterion validity addresses the extent to which test scores agree with an objective criterion that follows logically from the measured variable. The criterion that researchers select can be measured either at the same time as the variable to be validated or after that time.

Concurrent validity exists when a measure yields scores that are closely related to scores on a criterion measured at the same time. A store might validate a test of sales ability by administering it to sales personnel who are already employed and then comparing their test scores with their actual sales performance. Predictive validity is the ability of a measure to predict scores on a criterion measured in the future. For example, SAT scores might be predictive of college grade point averages.

Construct Validity

As we have learned, psychological constructs exist not as material entities but as ideas inferred on the basis of empirical relationships among measured variables. The question then becomes, how does one demonstrate the validity of a construct? Enter the multitrait–multimethod matrix, a table of correlations that is used to assess construct validity (Campbell & Fiske, 1959). Today, we view the multitrait–multimethod matrix as the gold standard for testing the validity of new ideas and new measurements in psychology. It tells you both how good your idea is and whether your idea is new and different.

In understanding the logic of a multitrait–multimethod approach, it is helpful to recall the advantages of using multiple methods of measurement. Now with a multitrait–multimethod matrix, you use not only two or more methods of measurement, but you also identify two or more theoretical constructs or traits to investigate. To simplify, let’s consider two traits, intelligence and extraversion; each is measured with two different methods of measurement: (1) behavioral observations and (2) tests of intelligence (e.g., WAIS-IV) and extraversion (e.g., NEO Five-Factor Personality Inventory; Costa & McCrae, 1985).

Within a single multitrait–multimethod matrix, two fundamental aspects of construct validity can be empirically evaluated. First is convergent validity, which is established by examining whether the same trait measured by two different methods yields similar results. In our example, intelligence measured by the WAIS-IV and behavioral observations should be related and so should extraversion measured by the NEO and behavioral observations. If different measures (e.g., behavioral observations, test scores) of the same trait correlate, then this would be considered strong empirical evidence for convergent validity. Second, the matrix provides a format to examine discriminant validity, which is demonstrated when different traits using the same form of measurement are unrelated. In our example, intelligence and extraversion should be unrelated, whether measured using tests or by behavioral observations. Thus, construct validity requires both convergent and discriminant validity, which can be most effectively empirically evaluated via the multitrait–multimethod matrix.

Exhibit 4.14 depicts a multitrait–multimethod matrix of our hypothetical example of two traits, intelligence and extraversion, and two methods of measurement, objective tests and behavioral observation. The psychometric methods of measurement are the WAIS-IV IQ test for the construct of intelligence and the NEO Five Factor Personality Inventory test for the construct of extraversion. There are three important points regarding the multitrait–multimethod matrix of correlational values presented in Exhibit 4.14. First, the correlational values presented in the diagonal (in parentheses) of the matrix are the Cronbach alpha reliability coefficients. This is a correlational statistic that indicates the degree of internal consistency for each of the four measurements; higher values reflect stronger reliability of the measure. As shown in the matrix, each of the four measurements had reliability coefficients of .90. Correlation values of this magnitude indicate highly reliable, internally consistent measures.
Second are the same-trait/different-method correlation coefficient values. Here, we see a .70 correlation coefficient between scores on the WAIS-IV IQ test and behavioral observation recordings for the construct of intelligence. We also see an identical .70 value for the correlation between scores on the NEO Five Factor Inventory test of extraversion and behavioral observation recordings for the construct of extraversion. Thus, the same-trait/different-method comparisons revealed high correlation values of .70 for intelligence and for personality. We therefore can conclude that our results demonstrated convergent validity for the each of these constructs of intelligence and extraversion. Third, the different-traits/same-method pairs for tests of intelligence and extraversion revealed a very low value of .30, as did the different-traits/same-method pairs for behavioral observations of intelligence and extraversion. For discriminant validity, there should be no relationship between two different constructs, such as intelligence and extraversion, that are tested with the same form of measurement, such as a psychometric instrument or behavioral observation.

Exhibit 4.15 presents the three important correlations generated from a multitrait–multimethod matrix. First is the reliability correlation of internal consistency of a trait measure (i.e., same trait/same method of measurement). Second is the convergent validity correlation of the same trait measured by two different methods, such as behavioral observations and test scores (i.e., same trait/different methods of measurement). A high correlation of two different methods of measurement of the same trait indicates high convergent validity. Third is the discriminant validity correlation of two different constructs measured by the same method. Here, a low correlation indicates high discriminant validity. Last are the nonsense correlations of different traits and different measures.
Latent Variable Models

Constructs can be inferred, developed, and evaluated through the use of a set of specialized statistical techniques that are referred to as **latent variable models** (Heinen, 1996). Perhaps the most well-known of these models is **factor analysis**, which can be used when a researcher wants to identify the distinct dimensions or “factors” of a given construct (DeCoster, 1998). For example, let us consider factor analyses of intelligence that were used in constructing the WAIS-IV. In this model, the construct of intelligence has been shown to encompass four distinct factors: (1) verbal comprehension, (2) perceptual reasoning, (3) working memory, and (4) processing speed.

How, you may ask, did this model come to be? Essentially, the model is a product of factor analyses, which may be viewed as a “bottom up” inductive procedure that investigates patterns of correlations among responses to a very large set of measurable indicators or variables. In this example, these indicators or variables are objective responses to an extensive sample of thousands of IQ test items. Factor analysis as a statistical technique uncovers what items coalesce or the extent to which participants respond similarly across various items. The more similar the scores, the more likely it is that this set of items taps a particular factor, such as working memory.

Note that the researcher labels a factor such as working memory based on what the interrelated items appear to be measuring. That is, the researcher infers the factor on the basis of the content of the interrelated items. Labeling and inferring factors are thus based on interpretation of the content that is measured. It is subjective, even though it is based on empirical findings, and as a result, it often needs to be confirmed on a second independent sample of participants given the same test items. This is referred to as **confirmatory factor analysis**, which is used to determine whether the same set of factors can be replicated in a new independent sample.

Constructs and factors are often used interchangeably. For our example, however, we used factor analysis to test the underlying structure of IQ. The four factors—working memory, processing speed, perceptual reasoning, and verbal comprehension—may be viewed as distinct dimensions of the construct of intelligence.

Culture and Measurement

Culture matters greatly in psychological measurement. Our measurements can be biased against certain cultures and ethnicities. For example, numerous studies have shown an **outcome bias** in IQ testing that places African Americans as a group at an approximately 15-point disadvantage in comparison with white groups (Neisser et al., 1996). Imagine flipping a coin that consistently comes up heads for any reason. In probability theory, the coin would be considered to be biased, regardless of any consequences that the outcome may or may not have. By analogy, an IQ test for which the black mean falls approximately 1 standard deviation (about 15 points) or less below the mean of whites would also be considered ipso facto evidence of outcome bias (Neisser et al., 1996).

In 1995, the social psychologists C. M. Steele and J. Aronson demonstrated how negative stereotypes (e.g., “genetic deficit”) that have been traditionally used to explain racial IQ differences detract from actual test performance of African Americans. This phenomenon is referred to as **stereotype threat**, and its deleterious effects on intelligence in African American students are shown in Exhibit 4.16. These data provide strong evidence that the racial disparity in IQ is due to environmental and social factors, not to a genetic deficit.

To understand how the disparity in IQ between racial groups has been erroneously interpreted as evidence of genetic inferiority, consider the following well-known plant analogy developed by the evolutionary biologist Richard Lewontin (1976), which is depicted in Exhibit 4.17. Imagine having a packet of virtually identical seeds, half of which you plant in a pot containing fertile soil and the other half in a pot of barren soil. The group of seeds planted in the poor soil will inevitably grow, but their growth will be stunted in comparison with the group of seeds planted in the rich soil.
**Exhibit 4.16** Stereotype Threat in IQ Performance in African Americans

**Effect of Stereotype Threat**

<table>
<thead>
<tr>
<th>Mean items solved (adjusted by SAT)</th>
<th>Stereotype Threat</th>
<th>No Stereotype Threat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blacks</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Whites</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

**Exhibit 4.17** The Plant Analogy Used to Explain Why Group Differences in IQ Are Determined by the Environment

- **Heritability = 100%**
  - Uniform lighting
  - Difference between groups is totally environmental

- **Uniform nutrient solution:** normal

- **Uniform nutrient solution:** deficient
Keep in mind that because the two groups of seeds are genetically identical, any differences in growth between the two potted plants are entirely due to the environment, which in this case is the soil. By the same token, within each group of seeds, some plants will grow larger than others. In contradistinction to the group differences observed between the two pots, the differences in plant heights within each pot are due to genetic variation within each group of seeds.

Now to apply this analogy to race/ethnic differences in IQ, we know that genetic variation can produce individual differences in IQ within a group (e.g., Bouchard & McGue, 1981), yet we also know that the average IQ difference between groups can still only be determined by the environment (e.g., Gray & Thompson, 2004). That is, genes along with environmental factors will influence individual differences in IQ, but only culture and environment will determine group differences in IQ. The critical lesson to take from this analogy is that any interpretation for racial/ethnic differences in IQ cannot use within-group genetic variation to account for between-group differences (Gray & Thompson, 2004). And the larger lesson for research methods in general is the importance of cultural considerations in any measurement in psychology (Na et al., 2010).

Conclusions

Researchers in psychology “play” with constructs. A construct is simply an idea, concept, or theoretical abstraction that has multiple referents, which in psychology are known as measurable variables. A construct is defined conceptually, a variable operationally. A theory unites a construct with its corresponding variables and their distinct forms of measurements. It does so by translating abstract concepts into particular measurable variables, each of which will include a form of objective measurement.

Research in psychology may thus be viewed as a dynamic interplay between conceptual and operational modes of thinking. On the one hand are theoretical constructs that exist as abstract formulations in need of empirical validation. On the other hand are operational definitions that act as recipes for specifying how constructs will be produced and measured. An operational definition specifies the scale of measurement that will be used to score a variable. A dependent variable can be scored by either one of four scales of measurement: nominal, ordinal, interval, or ratio. These scales of measurement are especially important when considering arithmetic calculations and statistical tests. Recall, for example, that a nominal scale reflects simple categorization for which any arithmetic calculations beyond basic counting are impossible. Knowing the scale of measurement of a variable is thus critical for what mathematical and statistical procedures are allowed to be computed. Measurement in the form of standardized tests is a field of study in and of itself and is known as psychometrics. Psychometrics is based on the principles of standardization of test construction, test administration, and test scoring.

A tool of measurement is judged by the extent to which it is a reliable and valid instrument. Reliability speaks to the consistency of a measurement, in terms of internal cohesiveness of its items and the stability of its scores over time. Validity speaks to whether the instrument measures what it is supposed to—e.g., the WAIS-IV measures intelligence, the SWLS measures subjective well-being, and so on. Questions of validity are not the same as questions of reliability: Reliability and validity are not synonymous. An instrument may be very reliable because it provides a highly consistent measure of a construct, but it may be invalid because it measures the wrong construct. The multitrait–multimethod matrix is a table of correlation coefficient values that can be calculated to compare and evaluate the reliability and validity of two or more psychological constructs. Last, measurement does not occur in a vacuum. The outcome bias of IQ testing underscores the importance of cultural considerations in any measurement in psychology.
Key Terms

Alternate-forms reliability
Concurrent validity
Confirmatory factor analysis
Construct
Construct validity
Content validity
Convergent validity
Criterion validity
Cronbach’s alpha
Dichotomy
Discriminant validity
Exhaustive
Face validity
Factor analysis
Interitem reliability
Interobserver reliability
Interval level of measurement
Latent variable models
Level of measurement
Likert scale
Measurement
Multitrait-multimethod matrix
Mutually exclusive
Nominal level of measurement
Nonparametric statistics
Normal curve
Norms
Ordinal level of measurement
Outcome bias
Parametric statistics
Positive psychology
Predictive validity
Psychological traits
Psychometrics
Random error of measurement
Ratio level of measurement
Reliability measure
Scale
Scatter diagram
Social desirability bias
Split-half reliability
Standard deviation
Standardization
Standardization sample
Stereotype threat
Stratified sample
Test-retest reliability

Research in the News

In a May 13, 2010, New York Times piece “Metric Mania,” the mathematics professor John Allen Paulos (2010a) writes, “The problem isn’t with statistical tests themselves but with what we do before and after we run them.” Let’s imagine that you have been assigned to write a similar piece for the New York Times, which you cleverly title “The Rule of Measurement.” In your piece, you want to explain to the reader that all statistics depend on reliable and valid measurements. That is, you want to explain why reliable and valid measurement is the lifeline of statistics as well as a vital component of all research. In so doing, describe the four different levels of measurement and how each can be applied to investigate a research question.

Activity Questions

1. Richard E. Nisbett (2009) argues in his book Intelligence and How to Get It: Why Schools and Culture Count that the disparity in average IQ between Americans of European and of African descent is “purely environmental.” Explain how
the Lewontin (1976) analogy shows that the racial disparities in IQ are best understood in terms of environmental factors.

2. Role-play what it would be like being a research participant assigned to either the stereotype threat or the nonstereotype threat condition of the Steele and Aronson (1995) study. Perhaps hand out exam blue books and read the two different sets of instructions to give your students a feel of what it would be like to be in either the stereotype threat or the nonstereotype threat condition.

3. The Implicit Association Test (IAT) is among the most widely researched web-based measures of covert thoughts and feelings. Go to https://implicit.harvard.edu/implicit/, and investigate the IAT, and take one of the IAT measures. Evaluate the IAT in terms of reliability and validity. What are the pros and cons of this relatively new technology for assessing covert thoughts and feelings? What are the cultural implications?

4. Are important constructs in psychological research always defined clearly? Are they defined consistently? Search the literature for six research articles that focus on “mental illness,” “mental health,” “happiness,” or some other construct suggested by your instructor. Is the construct defined clearly in each article? How similar are the definitions? Write up what you have found in a short report.

5. Two important human traits—intelligence and happiness—have attracted your curiosity. You are interested in construct validation of these two traits. You have at your disposal two methods of measurement: psychometric tests and behavioral observations. Show how you would use a multitrait–multimethod matrix to evaluate the construct validity of intelligence and happiness using these two different methods of measurement. Make sure to address both convergent validity and discriminant validity.

6. Imagine that you have been asked by a political consultant to devise a measure to assess whether people agree or disagree with a set of positions regarding a particular policy or political position. Describe how you would use a nominal level of measurement to gauge people’s opinions. What statistic would you use to summarize your hypothetical data that you plan to collect using a nominal level of measurement? What statistical test would you use to analyze your nominal data? Now describe how you would use an ordinal level of measurement to gauge people’s opinions. Again, identify the statistic you would use to summarize your ordinal data, and also indicate the best statistical approach for your data analysis. Finally, compare and contrast your nominal and ordinal levels of measurement.

7. Stereotype threat creates a significant problem in psychometric testing of cognitive abilities. Imagine that you wanted to study the effect of stereotype threat on high-level mathematics performance in women. Explain the stereotype threat and how it is theorized to detract from women’s performance in mathematics, especially at higher levels of complexity. What staged manipulation would you devise to create or operationalize stereotype threat in this study? How might stereotype threat be reduced so that people are judged by their abilities rather than by stereotype?

8. The construct of intelligence extends well beyond what is measured by IQ tests. For example, the psychologist Howard Gardner (1983, 1993) proposed the existence of multiple intelligences ranging from musical to logical-mathematical to interpersonal ability. What do you think intelligence means? What would your construct of intelligence entail? Would you include, as Howard Gardner does, the bodily kinesthetic intelligence that is shown by gifted athletes, surgeons, or dancers?

Review Questions

1. Think of a construct such as subjective well-being. How was it developed? How is it operationalized?
2. Why does a Likert scale use an ordinal level of measurement?
3. What are the advantages of using multiple methods or multiple forms of measurement such as behavioral observations and psychophysiological recordings?
4. Why is face validity inadequate? How would social desirability bias and face validity be related?
5. Why is IQ measured with an interval scale and income with a ratio scale?
6. Why can only the mode be used with data measured with a nominal scale? Why can’t the arithmetic mean and standard deviation be used with the nominal level of measurement?
7. What IQ score equals \( \frac{1}{3} \) standard deviation below the mean of 100?
8. What percentages of cases are expected to fall within ±1 standard deviation of the mean of the normal curve?
9. How would you stratify a sample of 2,000 people ranging in ages from 16 to 90 years?

10. How would you demonstrate that a test provides an internally consistent measure of a construct?

11. Why is discriminant validity important for construct validation?

12. Describe the relationship between the four levels of measurements and parametric/nonparametric statistics.

13. Describe the research design used in stereotype threat studies.

14. How does research on stereotype threat help explain the IQ disparity between Americans of European and of African descent?

15. Discuss the ethics of using staged manipulations in research.

16. Explain the principles of construct validity. Show how to construct a multitrait–multimethod matrix.

17. Why would alternate-forms reliability be important when comparing people’s memory before and after receiving a treatment drug?

18. What is outcome bias?

19. What kind of evidence would you cite to demonstrate the predictive validity of IQ?

---

**Ethics in Action**

1. What are the ethical implications of the erroneous conclusion that racial group differences in IQ are hardwired, fixed, and genetically immutable? How does such a mistaken claim lead to the ethically dubious argument that enrichment programs are of little or no benefit because of immutable group IQ differences?

2. Recall from Chapter 3 (“Research in the News”) the importance of fully informed consent for research participants with regard to how data collected from them in a given study will be used. Now consider informed consent for research participants in a study of racial differences in group IQ. What are the major ethical obligations a researcher has in ensuring that participants are fully informed as to the purpose of a study examining racial differences in group IQ?

3. In the stereotype threat studies by Steele and colleagues (1995) participants are given the same test, but some participants are essentially told that it is an IQ measure and other participants are told that it is a nondiagnostic measure of problem solving (i.e., not an IQ test). How would you justify this deception as ethical? How is it in keeping with the principle elucidated in the APA Ethical Code (see Chapter 3): *Avoid deception in research, except in limited circumstances*?