5

How Do We Measure Concepts?

Social researchers are faced with something of a dilemma when they attempt to measure concepts. Take, for example, a concept such as marital satisfaction. This concept is abstract; you cannot touch, feel, or see marital satisfaction. Yet, we somehow have to translate this abstraction into some kind of concrete **measurement.** We have to figure out a way to get people to tell us, as accurately as possible, how satisfied they are with their marriages.

Much of our thinking about families and children begins with what Babbie (2004) calls *conceptions.* We each have a conception of what defines marital satisfaction. But it's clear that marital satisfaction is not a physical object or quantity that can be readily observed and measured. How do we translate the relatively abstract and unobservable concept of marital satisfaction into something that we can measure and study?

Kaplan (1964) suggested three types of things that can be measured. A **direct observable** is a physical characteristic that can be observed directly, such as the number of people present at a meeting. A person's response to a questionnaire item about the number of people in a household is an **indirect observable**—an indirect representation of a characteristic or object. A **construct** is a creation based on observation, but it cannot be observed directly or indirectly. Marital satisfaction is an example of a construct. We may be able to observe consequences of marital satisfaction (e.g., marital stability), and we may be able to observe causes or antecedents of marital satisfaction (e.g., an equitable division of household labor between husband and wife); but we cannot observe marital satisfaction itself.

Operationalization is the name given to this process of translating abstractions into concrete measurement processes. In this process, we need to make a number of decisions about how to translate an abstract construct, such as marital satisfaction, so we can measure the marital satisfaction of our selected respondents.

Characteristics of a Good Measurement Procedure

In Chapter 2, I distinguished between variables and their states, levels, or attributes. Gender is a variable; female is an attribute of gender. Income is a variable; $45,000 is a state or level of income. One important decision to make in the operationalization process is what and how many attributes, states, or levels a variable will have in the measurement process.

A good measurement procedure is characterized by three key qualities. First, the set of attributes representing a variable must be exhaustive and inclusive. That is, all possible responses or states must be represented by a state, level, or attribute of the variable. If we were operationalizing the variable

marital status, we would find that using only the categories *married* and *single* would produce many blank stares from our respondents. We would probably need to include *never married, divorced but not remarried, married but separated,* and *widowed* as well.

Our second concern is that the attributes, states, or levels chosen to represent a variable must be mutually exclusive. No respondent should be able to be placed in more than one category. In the five-category scheme to measure marital status (never married, currently married, married but separated, divorced but not remarried, and widowed), no person could meaningfully claim to fall in more than one category.

Finally, we need to be concerned with our attribute scheme's precision. Ideally, our category scheme should use more, rather than fewer, distinctions when feasible. For example, rather than coding certain respondents as Protestant, it is probably more useful to include the choices Baptist, Episcopalian, Methodist, and so on. In this way you can determine how many respondents are of each denomination. But if you fail to allow respondents to describe whether they're Baptist, Episcopalian, or Methodist, you will never know.

Levels of Measurement

These three qualities of measurement processes—inclusive and exhaustive, mutually exclusive, and precise—are important to consider for all variables. The variables themselves, however, may be measured in vastly different ways. **Level of measurement** refers to the properties that define the measurement process itself. Most social and behavioral data are measured at one of four levels of measurement: nominal, ordinal, interval, or ratio.

Levels of Measurement: Nominal

Some variables have attributes that meet only the minimal requirements of being exhaustive and mutually exclusive. Some examples of nominal variables are religion, race, gender, city of residence, and marital status. Each of these variables, when properly defined, has a set of logically exhaustive and mutually exclusive categories or attributes. In other words, nominal variables identify the attributes associated with a variable without making any statements about whether one state or level is more or less than any other state or level. With variables measured at the nominal level, it's not possible to rank categories along any particular dimension, nor is it possible to state with great precision how much of the attribute is present.

It's also important to note that for nominally scaled variables, the differences between categories are not mathematically meaningful. For example,

let's say that we gave the following response choices to the question "What is your current marital status?"

1. Currently married
2. Married, but separated
3. Divorced, not remarried
4. Widowed
5. Never married

The fact that we assigned the number 5 to the response "never married" doesn't mean that it's one point better than "widowed" or five times as much as "currently married." For variables scaled at the nominal level, only the categories are meaningful—not the distances or gaps between categories or the numerical values we arbitrarily assign to the categories for coding purposes.

Levels of Measurement: Ordinal

Variables that are scaled at the **ordinal level of measurement** have all the properties of nominally scaled variables, but they add an important characteristic: The categories are ranked or ordered along some evaluative dimension. That is, unlike nominally scaled variables, ordinally scaled variables are assumed to represent some underlying continuum (e.g., from least to most or least desirable to most desirable). Some examples of variables that are typically measured at the ordinal level are social class (working class, middle class, upper class), marital happiness (not very happy, somewhat happy, very happy), or highest degree obtained (less than high school, high school diploma, college degree, graduate or professional degree).

Another type of measure that is often considered at the ordinal level is a **Likert scale** item (discussed further in Chapter 9). For example, we might ask respondents to indicate how they feel about this statement: "If a husband and wife both work full-time, they should share housework tasks equally." Do you

1. Strongly disagree
2. Disagree somewhat
3. Undecided, not sure
4. Agree somewhat
5. Strongly agree

Notice that such a variable has an exhaustive and mutually exclusive set of categories and that the categories clearly are ranked or ordered along the dimension of agreeing with the idea of sharing the housework. Those who say they "agree somewhat" with the statement are more strongly supportive of husbands and wives sharing housework than someone who says they "disagree somewhat." However, we can't say how much more supportive these people are, and we undoubtedly can't say that someone who agrees somewhat is twice as supportive of husbands and wives sharing housework as someone who disagrees somewhat. For ordinally scaled variables, the differences between categories are not mathematically meaningful because the numbers we assign to the categories merely reflect the ordering of the response choices.

One advantage of ordinal measurement is that people can often make comparative judgments relatively easily, whereas making an absolute judgment is far more difficult. For example, although respondents could probably tell us which aspect of their family lives gave them the most satisfaction, they would be hard pressed to tell us exactly how much more satisfaction they got from their family leisure time than from doing housework chores.

Levels of Measurement: Interval

Besides including all the characteristics of nominal and ordinal variables, variables that are measured at the **interval level** have differences between scale units that are meaningful and constant. Probably the most commonly used interval level measures employed in the social sciences are standardized intelligence tests. The scale is designed so that the 10-point difference between an intelligence quotient (IQ) score of 120 and one of 110 has the same meaning as the 10-point difference between scores of 140 and 130 (or, for that matter, any 10-point difference anywhere along the scale).

Although, strictly speaking, Likert-style items don't meet the requirements for interval-level measurement, they are often treated as if they do. Methodologists and statisticians don't seem to agree on whether this type of treatment is appropriate. One way to circumvent this criticism is to combine a number of Likert-style items into a composite index of some construct, and treat the scale total as an interval-level measure. There is more on this in Chapter 10.

Levels of Measurement: Ratio

Variables that are measured at the **ratio level of measurement** include all the characteristics of nominal, ordinal, and interval variables. What distinguishes

the ratio level of measurement from the other levels is that the scale of measurement has a theoretically defined origin, or zero point. Income in dollars is a ratio-level measure, because zero dollars means the absence of money. Number of children in a family is a ratio-level measure because zero children indicates the absence of children in the family. For comparison, note that temperature in kelvins is a ratio-level measure (because the zero point is theoretically defined as the absence of all molecular motion), but temperature in degrees Fahrenheit is not a ratio-level measure (because 0°F is an arbitrary point on the scale).

For ratio variables, as for interval variables, the scale intervals mean the same thing at all points on the scale (e.g., a $1 increase in income means the same at all points along the income continuum). For ratio variables, however, it also makes mathematical sense to compare different levels in a multiplicative sense. For example, a family with four children has twice as many children as one with two children; someone who earns $50,000 has twice as much income as someone who earns $25,000. But although 200 kelvins is twice as warm as 100 kelvins, an 80°F day is not twice as warm as a 40°F day.

Summary of Levels of Measurement

It may have occurred to you that the same concept might be measured at several different levels of measurement. Marital satisfaction, for example, can be measured at the nominal level by asking respondents, "Are you satisfied with your marriage?" and offering response choices of "Yes, I am satisfied with my marriage" and "No, I am not satisfied with my marriage." Marital satisfaction can be measured at the ordinal level by asking, "How satisfied are you with your marriage?" and offering choices of "very dissatisfied," "somewhat dissatisfied," "unsure," "somewhat satisfied," and "very satisfied." Similarly, family size might be measured at the ordinal level by classifying families as small, medium, and large or at the ratio level by finding out exactly how many members there are in a given family.

A key reason that level of measurement is so important is that different statistical techniques require measurement at particular levels. Some procedures, for example, are only appropriate for interval- or ratio-level data. Other techniques are appropriate for both nominal- and ordinal-level data, but some are specific to ordinal-level data. Whichever level of measurement is used, however, the key issues in data quality are reliability and validity.

What Are Reliability and Validity?

No measurement procedure is perfect. In statistical terms, no measurement procedure is error free. We know that some respondents will provide incorrect

responses because they don't know the answer to our question. Others will deliberately give wrong answers to appear more socially acceptable or answer the questions in the way they think the researcher wants them answered. Still other respondents will be confused by the phrasing or content of the question and answer incorrectly. Some respondents will give different answers to our questions at different times.

Social scientists gauge the amount of error in measurement in terms of reliability and validity. **Reliability** refers to the stability or consistency of a measurement operation. **Validity** refers to whether we're really measuring the concept that we intended to measure. More detailed discussions of each follow.

Reliability

If a measurement is really a good way of operationalizing a construct, that measurement procedure should yield the same result each time it is used (assuming, of course, that the value of the underlying construct hasn't changed over time). We refer to this type of reliability as **stability.** Another type of reliability deals with whether different observers using the same measurement procedure come up with the same result. This second type of reliability is known as **consistency.**

Stability

In the National Longitudinal Survey of Youth (NLSY), a sample of men and women were asked a series of questions about their work, education, and lives in general. In two separate interviews in 1983 and 1984, women were asked, "At what age did you first have sexual intercourse?" Obviously, a woman's age at first intercourse can't change over time. The response from any particular woman should not change from year to year if the measure is perfectly reliable.

If this measure were perfectly reliable, we would expect to get exactly the same response from each woman at each interview. Typically, however, some responses do change over time. Of the 5,788 women responding to this question at both interviews, only 57% gave the same response both times! Yet, of the women who gave different responses, nearly all (90%) gave a response in 1984 that was within two years of their 1983 response.

So, is this measure reliable or not? One way to think about this issue is to look at the difference between the average age given for first sexual intercourse at the 1983 response and compare it to the 1984 average. Overall, the average difference between the two responses is 0.03 year (about 11 days!). On that basis, we can conclude that the measure has very good reliability. Another way of looking at the reliability of this measure is to think in statistical terms. If this

measure were perfectly reliable, the correlation coefficient (a numerical index of similarity) between the 1983 and 1984 responses would be equal to 1.0. For these data, the correlation is about 0.97, suggesting high reliability in the sense of stability.

Consistency

Another way of thinking about reliability is in terms of consistency. Whereas stability refers to the reliability of a measure over time, consistency refers to its reliability across observers or across items. Reliability in the first sense is known as **interobserver reliability,** or agreement; reliability across items is known as **internal consistency,** or internal reliability. In Chapter 9, I discuss the issue of internal reliability in the context of index construction.

If we were doing a study of marital quality, we might have 6 family therapists observe 10 married couples discussing their marriages. We then ask each counselor to rate the quality of each couple's marriage on a scale of 1 (very poor) to 10 (very good). We would expect that if all 6 therapists are using the same criteria to judge the quality of the marriages, they would have a high degree of agreement. If the counselors agree on the marital quality for each couple, the measure of marital quality has reliability in the sense of consistency. We could even calculate an index of agreement where 0 equals no agreement whatsoever between the 6 therapists and 1.0 equals perfect agreement. Of course, just because 6 marriage counselors agree on how happy the couples are doesn't mean that their estimates are meaningful or valid.

Validity

In its truest sense, validity refers to how well we're measuring some underlying construct. Are we measuring what we think that we're measuring? The ideal way to measure the validity of a measurement operation is to compare the results of the measurement to the actual (real) value of the construct. Unfortunately, this usually isn't possible, because for most of the constructs we want to measure, actual value can't be determined. How can you learn a person's real attitude toward abortion? How could you know the real level of a person's marital satisfaction?

Consequently, most of our measures are indirect. We rely, for example, on a person's response to our questions about abortion to determine how that person feels about the abortion issue. When we attempt to get an indication of the real value of the construct, we are aware that our measurement process may be imperfect. Methodologists call a person's real attitude about abortion rights the *true* score; responses to our questions about abortion rights are the

observed score. Every observed score is the sum of two parts: the true score plus an error component.

The difference between the true score and the observed score is known as **measurement error.** Some of this error may be **random measurement error**: It is unrelated to the real value being measured. These random errors tend to cancel each other out and won't bias a measurement in any particular direction.

More serious for social and behavioral researchers is **systematic measurement error,** which can bias the result of the measurement process in one direction or another. For example, in a survey of parenting strategies, we might ask parents how many times in the past month they have had to physically punish their children. We expect to find that the responses are going to be biased downward—that is, we expect that parents systematically underreport the use of physical punishment—because many parents feel that use of physical punishment is indicative of a bad parent (even if they themselves use physical punishment). Moreover, parents who frequently use physical punishment are more likely to underreport such activity than those who rarely use physical punishment. Similarly, responses to a question on marital happiness might be systematically biased upward, because people often aren't willing to admit that they are unhappy with their marriages. In other words, the measurement error on these measures is probably not random. Singleton and Straits (2004) present a clear discussion of measurement error in the more general context of reliability and validity.

To deal with these problems, social scientists have developed a number of alternate indicators of a measure's validity.

Face Validity

When we talk about **face validity,** we mean whether the instrument looks like it's measuring what it's supposed to be measuring. A measure of marital satisfaction should not contain items about the legalization of marijuana or attitudes about gun control.

Of course, just because a measure of marital satisfaction contains items that *look like* they're measuring marital satisfaction doesn't necessarily mean that the instrument is measuring what it's supposed to be measuring. For example, we might ask a sample of married persons, "How satisfied are you with your marriage?" with response categories "not very satisfied," "somewhat satisfied," and "very satisfied." Such a measure looks valid from a face validity standpoint; clearly, the question is about marital satisfaction, and the response choices seem to capture the range of possible satisfaction levels. However, the item might have questionable validity if people tend to overreport their actual levels of marital satisfaction.

Criterion Validity

Criterion validity deals with how well a measure correlates with (or predicts) external criteria, including behaviors. There are two major forms of criterion validity. **Predictive validity** reflects the ability of a measure to predict future behaviors or conditions. We could expect our measure of marital satisfaction to correlate highly with how many violent arguments the couple had in the last 12 months. After all, you could assume that if a couple has violent arguments, they are not too satisfied with their marriage. Similarly, you would expect a valid measure of marital satisfaction to predict with some accuracy which couples would most likely divorce over the next five years.

Criterion validity can also reflect how well a measure correlates with alternative measures of the same phenomenon taken at approximately the same point of time; this type of criterion validity is known as **concurrent validity.** A new measure of spousal communication skills, for example, should correlate well with existing measures of such skills, especially those known to be well validated.

Construct Validity

Construct validity refers to how well a measure correlates with other theoretical constructs. We can expect a measure of marital satisfaction to correlate well with measures of such constructs as marital happiness, affection for one's spouse, and perceptions of fairness of the division of household tasks. If the measure of marital satisfaction did not correlate well with such measures, we would be suspicious that maybe it wasn't really measuring marital satisfaction.

Some writers argue that predictive or criterion validity is the most important indicator of a measure's overall validity. When measures of external criteria such as behavior are unavailable, however, construct validity can be a useful indication of how well we're measuring our concept.

Content Validity

Marital satisfaction probably has many different aspects or dimensions. Married persons might be satisfied or dissatisfied with their sex lives, how much work they have to do around the house, how much free time they have, how their paid work relates to their family life, how well they communicate with their spouses, and other dimensions as well. **Content validity** refers to how well the measure taps the full range of dimensions or meanings of the underlying construct.

Some Thoughts on Reliability and Validity

Imagine four different clocks. One of them is broken and always reads six o'clock. Another clock works perfectly but is set exactly 2 minutes ahead of the actual time. A third clock is fast and gains 1 minute every hour. The fourth clock keeps perfect time and is set correctly. Obviously, the fourth clock is both perfectly reliable and perfectly valid. But what about the other clocks?

The broken clock is perfectly reliable. No matter how many times we look at it, it always says the same time: six o'clock. What about the validity of this clock? The clock is exactly right—it is a perfectly valid measure of the current time—twice a day (at six o'clock in the morning and again at six o'clock at night). In fact, if we took our measurements at six o'clock each day, we might conclude that it is perfectly valid. Overall, however, this clock is not a valid measure of the current time, because it is correct only twice in each 24-hour period.

The second clock—the one that is always 2 minutes fast—is also perfectly reliable. If we check the clock each day at six o'clock, it will always read 6:02. As far as validity is concerned, this clock will never have the correct time, but it is an excellent indicator of the correct time.

The clock that gains 1 minute every hour is unreliable. If we check it at six o'clock each day, the clock's reading could be almost anything. Occasionally, this clock would have the correct time, but most of the time it's wrong.

The moral of this story is straightforward. A measure that is perfectly reliable is not necessarily perfectly valid. Yet, a measure that is perfectly valid is perfectly reliable; a measure that always gives the correct answer must always give the same answer (assuming that the underlying construct hasn't changed).

Sometimes a measure that we know is not perfectly valid may nonetheless be useful. None of our measures in the social and behavioral sciences are perfectly valid, but they don't have to be perfect to be of some value to us. Think about that the next time someone is critical of opinion poll results—while the numbers aren't perfect, they are probably a pretty good indication of how people feel.

Finally, keep in mind that the time to have these concerns about reliability and validity is before the instrument is constructed and before the data are collected. Once you've collected your data, it's too late to amend the research protocol. Whenever possible, it is important to assess the reliability and validity of the measures in our research *before* gathering our data. This may necessitate pilot studies or pretests where we gather a small amount of data and evaluate the reliability and validity of our measures before the full-scale data collection begins. It can be invaluable to simply ask someone not connected

with the research to read over the questionnaire or other instruments and make some general comments.

Study Questions

1. Give four examples of family-related variables measured at each of the four levels of measurement.
2. Select an article from a recent family journal (e.g., *Journal of Marriage and the Family, Journal of Family Issues,* or *Child Development*) and identify three of the major variables in the study. Then, for each variable, identify the level of measurement used in the study.
3. Using the same article you selected above, for each variable, discuss the reliability and validity concerns raised by the authors. If the authors don't make any statements about reliability and/or validity, make sure that you report this.

For Further Reading

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* San Francisco: Chandler.

Singleton, R., & Straits, B. C. (2004). *Approaches to social research* (4th ed.). New York: Oxford University Press.