4

The Art of Measurement

The first thing that must be understood about measurement is that nothing is measured directly. This is not just true in the social sciences; it is true also in the physical and biological sciences. To be sure, some measurements are more direct than others. But even our most conventional measures are not so direct. For example, we are so used to a thermometer measuring heat that we may forget that heat is an abstract theoretical concept that refers to the energy generated when molecules are moving. A thermometer reflects the principle that as molecules move more, a substance in a confined space (alcohol, mercury) will expand. We do not see increased heat; we see only the movement of the substance in the confined space (as when the mercury rises in a narrow cylinder). The number we get when we read a thermometer is not the same thing as the "heat" of a substance. And the number read from the thermometer is arbitrary. If we say the temperature is 54, that means nothing by itself. We use the number to see how the "heat" of a substance may change over time, or to compare the "heat" of one substance with the "heat" of another, or to relate "heat" to some other variable. The essential point of a measure (in this case a thermometer) is that it allows us to compare (cf. M. Ember 1970, 701).

But if all measurement is indirect, that doesn't mean that all measures are equally good. We are all familiar with inexpensive

bathroom scales that give you two different readings if you step on them twice. The trouble with the inexpensive scale is that it is not *reliable*. **Reliability**—consistency or stability—is one of the things we want from a measure. Inexpensive scales are not very *precise* either. That is, it is difficult to measure small fractions of a pound or a kilogram. But even the crudest measures can allow us to discover a pattern (e.g., the range of variation) or a relationship between one thing and another. For example, we don't have a very good way of measuring the height of infants because you have to stretch them out and they don't always cooperate. The ruler we measure them against may be precise, but straightening the baby's limbs is not. It doesn't really matter for most purposes. Medical practitioners can find out a lot about a baby's health and maturation by charting its height (however roughly measured) and weight at each visit and comparing these measures with averages for babies of the same age. Precision is nice, but not always necessary.

The most important requirement of a good measure is that it measures what it purports to measure. If a measure reflects or measures what it is supposed to, we say that it has **validity**. How do we know that a particular measure is valid if it, like all measures, is indirect? Trying to establish validity is one of the hardest things to do and we will discuss it in some detail later.

In order to evaluate a measure it is not just necessary to know the theoretical concept supposedly measured. It is also important to consider the cultural context and the purpose of the study. For example, two researchers might be interested in measuring economic productivity. One researcher might want to measure the average monetary equivalent of goods and services produced by the average adult in a day. But this measure would be meaningless in relatively noncommercial societies because much of production is not for sale; adults mostly produce food and other things for direct use by themselves and their families. A more widely applicable measure of productivity would reflect noncommercial as well as commercial production. Indeed, not all economic output can be measured in monetary terms, even in a society with buying and selling. If you want to measure

productivity, even in a commercial society, you have to try to measure the output of all production.

A few more general points about measurement. Be as explicit as possible about what concept you are trying to measure, and give others as much information as possible about how you measured that concept. A recipe is an appropriate analogy for the second part—giving others information. If you want a dish to come out the same way each time it is made, you have to try to spell out the ingredients and all the measurement and mixing procedures. (Of course, cookbook writers do not usually spell out their theoretical concepts; but we wouldn't think too much of a recipe that did not operationalize or spell out what you have to do!) Science depends upon replication. If we are to be confident about our findings, other researchers must be able to understand our measures well enough to be able to repeat them. Finally, since all measures are indirect (which is another way of saying that no measure is perfect), it is usually better to measure a concept in multiple ways.

To sum up what we have discussed so far:

- To measure is to compare one thing with another.
- All measures are indirect.
- Measures need to reflect the cultural context and purpose(s) of the study.
- Researchers should strive for reliability, precision, validity, and explicitness.
- Researchers should aim for multiple or alternative measures.

Designing Measures for Cross-Cultural Research Using Secondary Data

In chapter 2, we said that the research question suggests much of what the researcher needs to do. Suppose we asked a descriptive question: "What proportion of societies typically have extended families?" We know we have to measure the typical family form in a sample of societies. If we asked a causal

question, for example, "Why do some societies typically have extended families?" we would also have to measure the typical family form. However, in contrast to the descriptive question, the causal question hasn't told us what might explain the presence of such families. (The causes or predictors are not specified in the question.) Before we know what else to measure, we need to have at least one idea to test, one hypothesis about why such families exist.

Measures have to be specified for each variable (dependent and independent) in the hypothesis. The first steps involve the following:

1. Theoretically defining the variable of interest (in words or mathematically).
2. Operationally defining the measure(s), which means spelling out in practical terms the steps you need to take to make a decision about where a case falls on the "scale" that you have devised for measuring the variable. This is not always a straightforward process. Designing a measure requires some trial-and-error, and if the scale is too confusing or too hard to apply (because the required information is missing too often), the measure needs to be rethought.

Example: Extended Family Households

To illustrate the processes involved in measurement, let's start with a research question we have investigated (Pasternak, C. R. Ember, and M. Ember, 1976): Why do married couples in some societies typically live with related couples in extended family households, whereas in other societies couples typically live separately (in independent households)? This is a causal question that specifies the dependent variable, in this case, "type of household." By the way the question is phrased, the measure we design requires us to decide how to discover the typical household pattern in a society. The measure should be appropriate for a comparison of societies. Although the concept of extended family household may appear straightforward, we still

have to define it explicitly. The researcher should state what an "extended family" means, what a "household" means, and how to decide what is the typical type of household. The researcher may choose to define a family as a social and economic unit consisting minimally of one or more parents and children. An extended family might then be defined as consisting of two or more constituent families connected by a blood tie (most commonly a parent-child or brother-sister tie). And an extended family household might be defined as an extended family that lives co-residentially—in one house or in a compound or group of houses—and functions as an economic unit. An independent family would have only one constituent family. Having defined the concepts, the researcher must then specify the procedure for assessing what type of family is typical in the society. All of these steps are involved in operationalizing the variable of interest.

Definitions are not so hard to arrive at. What requires work is evaluating whether an operational definition is useful or easily applied. This is part of the art of measurement. For example, suppose we decided that in order to assess whether extended families or independent families were typical, we needed information from a census on the actual percentage of households of each type. From a measurement perspective this would be ideal. We could use as our measure the actual percentage of extended family households (0 to 100 percent). This would give us a **ratio measure** (see box 4.1). Other things being equal, a ratio measure (or at least an ordinal measure) is preferable to a nominal measure because we can use more powerful statistical tests with ratio or ordinal measures. (This is because such measures allow us to order our cases according to some scale.) However, pretesting would tell us that very few ethnographers provide the percentage of extended family households or give us the results of censuses. Rather, they usually say things like "extended family households are the norm." Or, "extended families are typical, but younger people are beginning to live in independent households." So our operational definition of percentage of extended family households, although ideal, may not be that useful, if we cannot find enough societies with reports based on household censuses.

Box 4.1. Types of Measurement

Nominal measurement. The simplest form of measurement divides things into discrete sets. The criteria for those sets should be clear and unambiguous so that we can classify a case as belonging to one or another set. If we decide to classify the typical household in a society as extended or independent, we are asking for a nominal classification into two sets. An example of a nominal measure with more than two sets is a form of marriage. Anthropologists typically classify marriages as monogamous, polygynous (one husband with two or more wives), or polyandrous (one wife with two or more husbands). While numbers may be assigned to these sets for entry into a computer, the numbers only convey that the cases are different.

While nominal scales are often referred to as the simplest form of measurement, they sometimes are the highest that can be meaningfully used. Take the concept of gender. If a society recognizes two genders of male and female, most researchers will simply want to classify people as one gender or the other.

Ordinal measurement. An ordinal measure adds an element of more or less to the concept of difference. If we line up individuals by their relative size we can give them rank order numbers that convey more or less tallness. If we give two people the same number, we convey that they are not different. Ordinal categories can be many, as in the lineup of the height of individuals, or they can be few such as "frequent," "occasional," and "rare." While numbers may be assigned to these relative positions, the numbers convey difference and order only.

Interval measurement and ratio measurement. If we wanted a measure of height we wouldn't normally line people up in rank order unless we didn't come equipped with a good long ruler. A good ruler has equally spaced intervals on a standardized metric. Instead of saying that one person is taller than another, we can say that they are taller by a certain number of centimeters or inches. With two numbers to compare on an interval or ratio scale, we can legitimately describe the *amount of difference* when we compare cases. From a statistical perspective, interval and ratio measures are essentially treated the same. Mathematically, however, a ratio scale has one additional property compared with an interval scale—it has a "true zero point." A thermometer calibrated on the Fahrenheit or Celsius scale is an interval scale, not a ratio scale because the zero point on both these scales does not mean the absence of heat. The Kelvin scale, on the other hand, is a ratio scale because "0" means that there is no motion of molecules and no "heat." An absolute zero point allows us to multiply and divide the numbers on the scale meaningfully. We can describe a person who is 6 feet high as twice as tall as a person who is 3 feet tall because the zero point on a ruler means "0" length. But it is not meaningful to say that when it is 60 degrees Fahrenheit it is twice as hot as when it is 30 degrees Fahrenheit.

Examples of interval and ratio scales:

population density
average rainfall in a year

population of the largest community
number of people in a polity
annual income
number of people in a household

Transforming one scale into another: All other things being equal, it is preferable to use a "higher" level of measurement. That is, an interval or ratio scale is preferable to an ordinal scale. An ordinal scale is preferable to a nominal scale. Leaving aside the concepts that are not meaningfully transformable, we can often imagine how we could transform a scale. As we saw in the extended family example, we can have a nominal scale that contrasts extended family household societies with independent family households. If we prefer, we can change the variable into an ordinal scale by relabeling the variable "frequency of extended family households" and we could classify societies as having frequencies that are "very high," "moderately high," "moderately low," or "infrequent or rare," according to the scale described in exercise 1. If we had enough information, we could employ a ratio scale with "percentage of households that are extended family households." What the researcher needs to be careful about is not to invoke a higher order of measurement or more precision when the data do not warrant it.

What can we do in this predicament? There are three choices:

1. We can stick to our insistence on the best measure and study only those societies for which the ethnography gives us percentages of each type of household (or percentages can be calculated from the quantitative information provided); we may have to expand our search (enlarge our sample) to find enough cases that have such precise information.
2. We can choose not to do the study because we can't measure the concept exactly the way we want to.
3. We can redesign our measure to incorporate descriptions in words that are not based on census materials.

Faced with these three choices, most cross-cultural researchers would opt to redesign the measure to incorporate word descriptions. (That would be our choice.) Word descriptions

do convey information about degree, even if not as precisely as percentages. If an ethnographer says "extended family households are typical," we do not know if that means 50 percent or 100 percent, but we can be very confident it does not mean 0 to 40 percent. And we can be fairly sure it does not mean 40 to 49 percent. If the relative frequency of extended family households (measured on the basis of words) is related to something else, we should be able to see the relationship even though we are not able to use a percentage measure based on numerical information. Relative frequency is a type of **ordinal measure**, where numbers on the scale convey an ordering from more to less. A measure of relative frequency of extended family households might read something like what follows.

Code extended family household as:

4. *Very high* in frequency if the ethnographer describes this type of household as the norm or typical in the absence of any indication of another common type of household. Phrases like "almost all households are extended" are clear indicators. *Do not use discussions of the "ideal" household to measure relative frequency, unless there are indications that the ideal is also practiced. If there is a developmental cycle, such as the household splitting up when the third generation reaches a certain age, do not use this category. Rather, you should use scale score 3 if the extended family household remains together for a substantial portion of the life cycle or scale score 2 if the household remains together only briefly.*
3. *Moderately high* in frequency if the ethnographer describes another fairly frequent household pattern but indicates that extended family households are still the most common.
2. *Moderately low* in frequency if the ethnographer describes extended family households as alternative or a second choice (another form of household is said to be typical).
1. *Infrequent or rare* if another form of household is the only form of household mentioned and if the extended family form is mentioned as absent or an unusual situation.

Do not infer the absence of extended families merely from the absence of discussion of family and household type. To use this category, the ethnographer must explicitly discuss family and household.

Don't know if there is no information on form of household, or there is contradictory information. (We usually use discontinuous numbers like 8 for no information and 9 for contradictory information; these numbers need to be dropped before performing statistical tests. The discontinuity helps remind the researcher that these numbers should not be included.)

It is very important to tell your coders how to infer that something is rare or absent. Most ethnographers do not give an inventory of what is missing in a society. Researchers therefore have to specify the appropriate rules for inferring rarity or absence. In scale point 1 above, our rules specify that the coder is not allowed to say extended families are absent in the absence of information. If there is no information the coder must say "don't know." We will return to this issue later.

A final part of the instructions is to specify how to find the information required to make a decision. Many cross-cultural researchers use the Human Relations Area Files (HRAF) Collection of Ethnography (see the description in chapter 6 and the appendix on the HRAF Collections). This full-text database is complexly subject-indexed, so that a researcher can rapidly find paragraphs relevant to the indexed subject. (Word searches are also possible in the electronic HRAF, *eHRAF World Cultures*, or you could search using a combination of subject categories and words in the text.) It is very easy to find the relevant information when the subject matter of interest to the research is clearly indexed in the HRAF Collections. Regarding our concern here, the HRAF Collection of Ethnography has one subject category (592, Household) that can be consulted. Another advantage of this database is that the independent and dependent variables can be measured in separate steps, which minimizes the chance that knowing one variable will influence the coding of another variable.

Notice the italicized caveats in the above scale on extended family households. These caveats are usually inserted after the researcher realizes the practical problems that may arise when looking at actual ethnographies (this stage is what we call *pretesting*). Additional *pretesting* should be done using coders who have not had anything to do with creating the scale. It may turn out that four distinctions are too difficult to apply to the word descriptions usually found in ethnographies, so a researcher might want to collapse the scale a little. Or, it may turn out that two coders do not frequently agree with each other. If so, the investigator may have to spell out the rules a little more. And if we use the *ordinal* scale described above, what do we do when the ethnography actually gives us precise numbers or percentages for a case? It is usually easy to fit those numbers into the word scale (or to average two adjacent scale scores). So, for instance, if 70 percent of the households have extended families, and 30 percent are independent, we would choose scale score 3. But we might decide to use two scales: a more precise one based on numerical measurement (percentages) for those cases with numbers or percentages, the other scale relying on words (when the ethnography provides only words). C. R. Ember et al. (1991) recommend the use of both types of scale when possible. The advantage of using two scales of varying precision is that the more precise one (the quantitative scale) should be more strongly related to other variables than the less precise scale. (The less precise scale should be less accurate than the more precise scale, assuming that the less precise one has to rely sometimes on ambiguous words.) Stronger results with the more precise scale would increase our confidence that the relationship observed even with the less precise scale is true.

Alternative to New Measures: Using Existing Measures

The measurement scale described above is a hypothetical one. In our study of extended family households (Pasternak, C. R. Ember, and M. Ember 1976) we actually decided to use an existing measure of extended family households that was used in the

Ethnographic Atlas (Murdock 1967) for our dependent variable. We had a lot of discussion about the wisdom of this choice. Our main reason for deciding to use an existing measure was that we knew we had to code the independent variable ourselves. As far as we were aware, "incompatible activity requirements," our independent variable, had never been measured before. (Incompatible activity requirements refers to the need for a given gender to do two different tasks at the same time in different places, such as agricultural work and childtending; we reasoned that with such requirements, two or more individuals of the same gender would be needed in the household—hence extended family households would be favored.) We thought it wiser not to code both the independent and dependent variables ourselves. After all, it was our theory we were testing. One of the maxims in science is to try to minimize unconscious biasing of the results. So we decided to measure incompatibility requirements ourselves and to create a score on extended family households using ratings from the *Ethnographic Atlas* (Murdock 1967, col. 14) on extended family households. An additional advantage to using Murdock's measure is time. If we coded both the independent and dependent variables ourselves, it would have taken us a lot longer to measure our variables.

The drawback to using a previous researcher's measure is that it may be difficult to be confident that the previous measure is actually measuring what you want to be measuring. *The most serious mistake is to use an existing measure that is not really what you want conceptually.* As we discuss below, lack of fit between the theoretical concept and the operational measure is a serious source of error. In our case, we could be fairly confident about Murdock's measure of type of household. From what we had read of his writings on social structure, his definitions of family and household were similar to ours. (He was Melvin Ember's mentor in graduate school at Yale.)

If you do decide to use someone else's measure, your operational measure becomes a description of how you used the other researcher's scale. If the other researcher's definitions are published in an accessible place, you can refer the reader to them

without repeating all the definitions. For instance, we were able to say the following:

> Extended family households were considered to be present if the Atlas said the case typically had extended family households (E, F, or G in Column 14) and were considered to be absent if the *Atlas* said the case typically had independent families (M, N, O, P, Q, R, or S in Column 14). By extended family household, the *Atlas* means that at least two related families, disregarding polygamous unions, comprise the household—which is equivalent to our concern in this study. (Pasternak, C. R. Ember, and M. Ember 1976, 119)

As it happened, incompatible activity requirements (as measured by us) strongly predicted extended family households (as measured by Murdock 1967). But what if the results hadn't worked out so well? We might then have concluded that our theory was questionable. But it is also possible to get falsifying results if you have too much measurement error. At that point, we might have decided to measure extended family households on our own, using a new ordinal scale, to see if the results would improve.

Minimizing Error in the Design of Measures in Secondary Comparisons

Measures are designed to tap theoretical concepts. Ideally we want the measure of a concept to be free of error. If a measure taps the concept exactly, it would be a valid measure. The more the measured score departs from the theoretical construct, the less valid is the measure. In secondary comparisons, error can come from a variety of sources. There may be errors by the ethnographer or the original informants, and there may be errors by those reading and coding the ethnographic information for measures. These types of errors and the steps that can be taken to minimize them will be addressed in the next chapter. Here we address the errors that may result from the lack of fit between

the theoretical concept and the designed measure. If the designed measure is measuring something other than the theoretical construct, the researcher is building in serious error, which no amount of carefulness in other aspects of the research design can undo. For the whole point of deriving a hypothesis from a theory is to test that theory. If the measures of the concepts in the theory are far removed from the constructs—if the measures are not valid—it is not legitimate to claim that the tests using those measures can allow us to evaluate the theory.

Types of Validity Used in Secondary Cross-Cultural Research

How can validity be established? The dilemma of all research is that theoretical constructs cannot be measured directly. Therefore there is never any certainty that a measure measures what it is supposed to measure. Even though all measurement is indirect and therefore validity cannot ever be established beyond all doubt (Campbell 1988), some measures are more direct and are therefore more likely to be valid than others. So, for example, more direct measures arouse little doubt that they are measuring what they are supposed to measure. They have high **face validity**; there is little or no need to justify why we are confident about their validity. Other things being equal, we suggest that cross-culturalists try to use measures that are as direct as possible, because less inference and less guesswork generally yield more accuracy and hence stronger results (assuming that you are dealing with a true relationship). For example, when a cross-culturalist wants to measure whether a husband and wife work together, it is more direct to use a measure that is based on explicit ethnographers' reports of work patterns than to infer the work pattern from general statements about how husbands and wives get along (C. R. Ember et al. 1991, 193). A measure based on how well husbands and wives get along would have low face validity as a measure of husbands and wives working together, but a measure based on ethnographer reports of work patterns would have high face validity.

Let's consider the hypothetical measure we constructed (described above) for the frequency of extended family households. The measure we ourselves designed would require the coder to read ethnographers' statements about family type and use the ethnographers' words about frequency to judge the prevalence of extended family households. The measure is very close in meaning to the theoretical construct and on the face of it seems valid.

While direct measures with high face validity are preferable, sometimes a researcher wants to measure something more difficult. The theoretical construct may be quite abstract. Consider the following constructs: community integration, emotional expressiveness, the status of women, or cultural complexity. It is hard to imagine a clear, direct measure of any of these constructs. Rather, we might imagine a lot of different indicators. For example, with respect to the status of women, we may imagine that it could be reflected in leadership positions, in decision making in the household, in the gender and power of the gods, and so on. Similarly, cultural complexity may be indicated by subsistence technology, the number of different specialists, the degree to which there is a hierarchy of political authorities, and so on.

Other types of validation techniques are harder to use in cross-cultural research using secondary data, but they may be useful in comparisons of field data. These validation techniques involve showing that the new measure is highly correlated with another generally accepted measure (the criterion). The criterion variable may pertain to the future (such as future school performance), the past (prior school performance), or it may be another measure pertaining to roughly the same time.

When no clearly accepted measure is available as the criterion, a measure may be judged in terms of **content validity**—the degree to which "a specified domain of content is sampled" (Nunnally 1978, 91). If ability in some subject is to be measured, a test that covers a lot of different domains would probably be more valid than a test covering only a few selected domains. In cross-cultural studies of abstract constructs, it may be a good idea as you begin to develop your measures to measure items

across a wide variety of domains. So, for example, in order to assess the status of women (an abstract construct), Whyte (1978) measured fifty-two different variables that might tap the relative status of women in a wide array of domains. Some of those domains were family life and decision making, economic roles and control of resources, political leadership, and religious beliefs and practices. Broude and Greene (1983), looking at husband-wife intimacy, measured patterns of eating, sleeping, leisure-time activities, work relations, and husbands' attendance at the birth of a child. Perhaps the most widespread use of content validity is with regard to measures of cultural complexity. For example, Carneiro (Carneiro and Tobias 1963; Carneiro 1970) measured as many as 618 traits covering a broad range of domains presumably relating to complexity. Other measures of cultural complexity use fewer traits, but most span a broad array of domains. The presumption behind the concept of *content validity* is that a measure is more likely to be valid if it taps into all or most of the relevant domains.

Does this mean that the more traits (or items) included, the better the measure? Not necessarily. First, items that do not tap the same dimension of variation do not improve a measure. It is possible to discover whether items belong together in a given scale (see Weller 1998, and Handwerker and Borgatti 1998, for an introduction to the voluminous literature on methods of scaling). Second, there are practical issues to consider as well as theoretical ones. Too many items can make the research impractical. In conducting a personal interview, too many questions can lead to fatigue (on the part of the researcher as well as the interviewee). In cross-cultural research using ethnographic materials, measuring even a simple trait could take thirty to sixty minutes of reading per case, in our experience. So measuring hundreds of traits would be a formidable task and might lead one to sacrifice other principles of good research design, such as the desirability of studying a random sample of cases that was large enough to allow for statistically significant results even with a lot of missing data.

Statistical techniques can be used to test whether a set of items belongs together. Such tests can be done on small samples

to pare down the list of items that need to be used on larger samples. In scaling items or traits that may or may not belong together, there are some general principles.

First, if all the traits or items employed tap the same underlying concept (dimension), they should all be associated with each other, at least to a moderate degree. If you have several measures of baby care—dealing separately with how much the mother holds, feeds, and plays with a baby—they may correlate with each other. But they may not. The researcher may find that various traits have a more complicated or multidimensional structure. For example, in many societies a sibling carries the infant around most of the time, and brings the baby to the mother only for nursing. Thus, it may turn out that not all aspects of baby care cluster together along one dimension of more or less attentiveness. It may be necessary to use two or more scales to tap the multiple dimensions of baby care. Second, if not all the traits or items are indicators of the same construct, it is nearly always possible to discover how many dimensions account for their similarities and differences (Weller 1998; Handwerker and Borgatti 1998).

In the case of cultural complexity, we have scales that use many or only a few traits, from more than six hundred (Carneiro 1970, 854–70) to three (Naroll 1956) to two (Marsh 1967). Researchers developing new scales try to compare them with previously developed scales. As it turns out, all of these scales are highly correlated with each other (Peregrine, Ember, and Ember 2000). The concept of **convergent validity** (Campbell and Fiske 1959) refers to the situation of strong intercorrelations among a number of independent measures. If a number of different scales are strongly related, researchers could decide to use the scale or scales that are easier to apply or that best fit the theoretical purposes of their research.

Suppose a researcher is interested in general trends in cultural evolution. In this case, it makes the most sense to use a measure that is designed to tap a sequence of development. For example, one could try to use a **Guttman scale**. This kind of scale is hierarchical. The items are listed in an evolutionary order. If a case has the highest item on the scale, it is likely to have all the

other items. If it has a score halfway down, it is likely to have half the items. This steplike feature holds for scores anywhere on the scale. Each case is scored as having either the trait present (in which case it receives a 1 for that trait) or the trait absent (in which case it receives a 0 for that trait). Linton Freeman (1957 as described in Tatje and Naroll 1970) found that the following items scaled in the hierarchical order shown below:

1. Presence or absence of trade with other societies.
2. Presence or absence of a subsistence economy based primarily on agriculture or pastoralism.
3. Presence or absence of social stratification or slavery.
4. Presence or absence of full-time governmental specialists.
5. Presence or absence of full-time religious or magical specialists.
6. Presence or absence of secondary tools (tools fashioned exclusively for the manufacture of other tools).
7. Presence or absence of full-time craft specialists.
8. Presence or absence of a standard medium of exchange with a value fixed at some worth other than its commodity value.
9. Presence or absence of a state of at least ten thousand in population.
10. Presence or absence of towns exceeding one thousand in population.
11. Presence or absence of a complex, unambiguously written language.

If a case receives a score of 11 on this scale, it means two things. One is that it has a complex, written language. The second thing implied by a score of 11 is that the case is likely to have all the other items. That is, each score implies the presence also of the traits marked by lower numbers. Notice that the scale is hierarchical only in one direction; the presence of a town of more than one thousand population (a score of 10) does not imply a written language (a score of 11), but it does imply the presence of item 9 (a state of at least ten thousand population). Note too

that the other traits lower in the scale may not be present with certainty, but if the scale is a Guttman-type scale, it is very likely that the traits lower are present. Most cases with a scale score of 6, for example, would also have items 1 through 5.

Comparative Projects Collecting Primary Data

Since this book focuses primarily on secondary cross-cultural comparisons (using other people's data), we will only briefly discuss measurement in the field. Other methods sources should be consulted for detailed discussions on collecting primary data in the field and judging informant accuracy (see, for example, Romney et al. 1986; Bernard 1994; Johnson and Sackett 1998; Weller 1998). And see the Munroes (1991a, 1991b) and Johnson (1991) for extended discussions of comparative field studies. Some measurement techniques lend themselves more readily to comparison than others. While participant observation and unstructured interviewing are usually a necessary first step in any field study, such techniques are less likely to lead to comparable data-gathering and measurement across sites than more structured interviews and systematic observation. It goes without saying that any measure used across sites needs to be applicable to all the sites. Most comparative field studies deal with general domains of life that are found in all societies—for example, words and classification of colors, kin, animals, and plants, ideas about illness, raising children, and social behaviors.

Comparing Primary Data from Field Studies: Behavior Observations

A comparative researcher may be interested in kinds of data that are very unlikely to be described in the ethnographic record. So conventional cross-cultural comparison, using the secondary data in the ethnographic record, is not possible. If you can't find enough information in ethnographies to measure the variables of interest to you, what else can you do? One possibility is to do a comparative field study (R. L. Munroe and R. H. Munroe

1991a, 1991b). If there is little relevant ethnography available, you could collect behavior observations systematically in more than one field site on the variables of interest.

But, as Beatrice and John Whiting (1970, 284) pointed out, systematic behavior observation is so time-consuming that comparative field workers should only consider it if the interviewing of informants cannot provide valid data on the domain in question. For example, adults (particularly males) may not be able to report accurately on child rearing or child behavior. Or ethnographers (generally men, in the early years of anthropology) may not have been interested generally in collecting such information. Either way, there were few extensive descriptions of child rearing in the ethnographic record as of the 1950s. The Six Cultures Project (J. W. M. Whiting et al. 1966; B. B. Whiting and J. W. M. Whiting 1975) and the Four Cultures Project directed by Robert L. and Ruth H. Munroe (R. L. Munroe et al. 2000; see references therein) were designed as comparative field projects to measure children's behavior. To illustrate how you can design measures for comparative field studies, we focus here on how children's aggressive behavior was measured by the Munroes.

The investigators (R. L. Munroe et al. 2000, 7) code behavior observations following the classification suggested by the Whitings (B. B. Whiting and J. W. M. Whiting 1975, 54–65). The Whitings classify social behaviors into twelve major categories. The Munroes consider three of them to involve "aggressive behavior" (2000, 7)—assault, horseplay, and symbolic aggression. The Whitings (1975, 57–62) suggest that the observer consider whether the observed acts of physical aggression were serious or playful. Serious assaults included such behaviors as striking, slapping, or kicking someone else; playful or sociable assaults consisted of behaviors such as friendly wrestling or backslapping. Symbolic aggression included threats or insults by word or gesture as well as attempts to frighten or derogate another person. Responsible aggression—for example, aggression that was legitimately administered as a punishment—was excluded by the Whitings from the aggressive categories.

The Munroes (R. L. Munroe et al. 2000) trained local native assistants to observe the social behavior of children and record

the behavior according to the twelve major behavior categories specified by the Whitings. The Munroes planned to construct a proportion score for each type of aggression observed (the number of each type of aggression divided by the total number of social acts observed). Each observer was assigned children to observe; on a given day, the observer would look for a particular (focal) child in his or her natural setting. If the focal child was interacting with at least two other persons, the observer was told to watch for any social behavior on the part of the focal child (either the child initiating a social interaction with someone else or the child responding to someone else). The first observed social behavior was described. The descriptive protocol included the details of the behavior, the type of behavior that the child or another individual may have initiated (which of the twelve categories the behavior belonged to), the type of response to the initiation, and the individuals within close proximity to the focal child. In contrast to the Whitings' Six Cultures Project, in which the observers recorded behaviors for five minutes, the Munroes decided to record only the first social behavior observed. With the measures they constructed, the Munroes demonstrated significant overall differences by gender—male children were generally more aggressive, in all categories of aggression but especially in assault. In addition, aggression generally declines with age and is much more likely the more male peers are present in the social setting. The Munroes' findings from the four field sites are generally consistent with previous research in other cultures employing somewhat different measurement techniques. For example, the Six Cultures Project (J. W. M. Whiting et al. 1966; B. B. Whiting and J. W. M. Whiting 1975) used longer behavior protocols (five minutes in length). Robert Munroe (personal communication) suggests that the short observations the Munroes employed (in which the observer only records the first social behavior seen) are less intrusive and allow the collection of observations over a larger span of time. The disadvantage is that the observer is not catching sequences of behaviors.

The major obstacle to choosing to use systematic behavior observations in comparative field studies is that they take a long time to do (you have to spend considerable time in the field, in

more than one place) and they are very expensive (not only must the investigators be supported, but also the native observers). So, as much as most researchers would like to obtain data directly, by comparative field studies, conventional cross-cultural research using the data already in the ethnographic record is much more feasible. But remember: you can't do a conventional type of study if you can't think of how you can use data in the ethnographic record to measure the variables of interest to you.