

Handbook of Methods in Cultural Anthropology

Copyrighted Material
Not for Reproduction

Copyrighted Material
Not for Reproduction

Handbook of Methods in Cultural Anthropology

Second Edition

EDITED BY

H. RUSSELL BERNARD AND CLARENCE C. GRAVLEE

Copyrighted Material
Not for Reproduction

ROWMAN & LITTLEFIELD

Lanham • Boulder • New York • London

Published by Rowman & Littlefield
A wholly owned subsidiary of The Rowman & Littlefield Publishing Group, Inc.
4501 Forbes Boulevard, Suite 200, Lanham, Maryland 20706
www.rowman.com

16 Carlisle Street, London W1D 3 BT, United Kingdom

Copyright © 2015 by Rowman & Littlefield
First edition copyright © 1998 by AltaMira Press

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without written permission from the publisher, except by a reviewer who may quote passages in a review.

British Library Cataloguing in Publication Information Available

Library of Congress Cataloging-in-Publication Data

Handbook of methods in cultural anthropology / edited by H. Russell Bernard and Clarence C. Gravlee. — Second edition.

pages cm

Includes bibliographical references and index.


ISBN 978-0-7591-2070-9 (cloth : alk. paper) — ISBN 978-0-7591-2071-6 (pbk. : alk. paper)

— ISBN 978-0-7591-2072-3 (electronic) 1. Ethnology—Methodology. I. Bernard, H. Russell (Harvey Russell), 1940–

GN345.H37 2015

305.8001—dc23

2014007881

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI/NISO Z39.48-1992.

Printed in the United States of America

Contents

Preface	vii
Introduction: On Method and Methods in Anthropology <i>H. Russell Bernard and Clarence C. Gravlee</i>	1
PART I. PERSPECTIVES	
1 Epistemology: The Nature and Validation of Knowledge <i>Michael Schnegg</i>	21
2 In Search of Meaningful Methods <i>James W. Fernandez and Michael Herzfeld</i>	55
3 Research Design and Research Strategies <i>Jeffrey C. Johnson and Daniel J. Hruschka</i>	97
4 Ethics <i>Carolyn Fluehr-Lobban</i>	131
5 Feminist Methods <i>Christine Ward Gailey</i>	151
6 Participatory Methods and Community-Based Collaborations <i>Stephen L. Schensul, Jean J. Schensul, Merrill Singer, Margaret Weeks, and Marie Brault</i>	185
PART II. ACQUIRING INFORMATION	
7 Sampling and Selecting Participants in Field Research <i>Greg Guest</i>	215
8 Participant Observation <i>Kathleen Musante (DeWalt)</i>	251
9 Behavioral Observation <i>Raymond Hames and Michael Paolisso</i>	293
10 Person-Centered Interviewing and Observation <i>Robert I. Levy and Douglas W. Hollan</i>	313

11	Structured Interviewing and Questionnaire Construction <i>Susan C. Weller</i>	343
12	Discourse-Centered Methods <i>Brenda Farnell and Laura R. Graham</i>	391
13	Visual Anthropology <i>Fadwa El Guindi</i>	439
14	Ethnography of Online Cultures <i>Jeffrey G. Snodgrass</i>	465
15	Social Survey Methods <i>William W. Dressler and Kathryn S. Oths</i>	497

PART III. INTERPRETING INFORMATION

16	Reasoning with Numbers <i>W. Penn Handwerker and Stephen P. Borgatti</i>	519
17	Text Analysis <i>Amber Wutich, Gery Ryan, and H. Russell Bernard</i>	533
18	Cross-Cultural Research <i>Carol R. Ember, Melvin Ember, and Peter N. Peregrine</i>	561
19	Geospatial Analysis <i>Eduardo S. Brondizio and Tracy Van Holt</i>	601
20	Social Network Analysis <i>Christopher McCarty and José Luis Molina</i>	631

PART IV. APPLYING AND PRESENTING INFORMATION

21	Theories and Methods in Applied Anthropology <i>Robert T. Trotter, II, Jean J. Schensul, and Kristin M. Kostick</i>	661
22	Presenting Anthropology to Diverse Audiences <i>Conrad Phillip Kottak</i>	695
23	Public Anthropology <i>Thomas Hylland Eriksen</i>	719
	Author Index	735
	Subject Index	763

Cross-Cultural Research

CAROL R. EMBER, MELVIN EMBER,
AND PETER N. PEREGRINE

CULTURAL ANTHROPOLOGY, AMONG OTHER THINGS, IS A COMPARATIVE DISCIPLINE. We who call ourselves cultural anthropologists like to emphasize how customs vary from place to place and how they may change over time. Indeed, we delight in the diversity of human cultures. Yet many cultural anthropologists are uncomfortable with the idea of explicit, systematic, cross-cultural comparison—the subject of this chapter. One reason for the discomfort may be our emphasis on fieldwork. We train several years for our fieldwork and spend a lot of time in the field. Fieldwork is central to our professional lives as well as to the discipline. The usual objective of fieldwork is to discover the details and particulars of a single community or culture. Those details remind us that each culture is unique, its combination of patterns of behavior and belief like no other.

To compare cultures is not to deny their individual uniqueness. Ethnography tells us what is distinctive about a particular culture; cross-cultural comparison tells us about what is generally true for some, many, or even all human cultures. To generalize across cultures, we build on the particulars of ethnographies to formulate statements about the similarities and differences of cultures and what they may be related to. The serious epistemological issue is whether it is possible to formulate such general statements in the first place. Cross-culturalists argue that it is.

We focus in this chapter on methods for systematic comparisons across cultures—comparisons that, we expect, will answer questions about the incidence, distribution, and causes of cultural variation (see C. R. Ember and M. Ember 2009 and Special Issue 1991). The methods are familiar—unbiased sampling, repeatable measurements, statistical evaluation of results, and the like. The relationship between cross-cultural research and ethnography is analogous to that between epidemiology and clinical practice in medicine. In ethnographic research and in clinical practice, the focus is on the individual case, while in cross-cultural research, as in epidemiology, the focus is on populations. Epidemiologists look at the incidence and distribution of diseases across populations and try to understand the causes of those diseases, primarily through correlational analyses of presumed causes and effects. Similarly, cross-cultural researchers are interested in causes and effects of cultural variation across the world or across regions of the world.

UNIQUENESS AND COMPARABILITY

To illustrate how things can be unique and comparable at the same time, consider the following ethnographic statements about sexuality in three different cultures:

1. The Mae Enga in the Western Highlands of Papua New Guinea believe that “copulation is in itself detrimental to male well-being. Men believe that the vital fluid residing in a man’s skin makes it sound and handsome, a condition that determines and reflects his mental vigor and self-confidence. This fluid also manifests itself as his semen. Hence, every ejaculation depletes his vitality, and over-indulgence must dull his mind and leave his body permanently exhausted and withered” (Meggitt 1964, 210).
2. “The Nupe men [of Nigeria], certainly, make much of the physically weakening effects of sexual intercourse, and teach the younger generation to husband their strength” (Nadel 1954, 179).
3. “[T]he milk avoidances of the Nilotes [Shilluk of the Sudan] are dependent on fear of contamination associated with the sexual act. . . . Only small boys herd the cattle and milk them, for once a boy has reached maturity there is the danger that he may have had sexual contact, when if he milked, or handled manure, or even walked among the cattle in their pens, he would cause them to become sterile. . . . If a man has had sexual relations with his wife or another he is considered unclean and does not drink milk until the sun has set the following day” (Seligman and Seligman 1932, 73).

Each statement about male sexuality is unique, but there are also similarities in these statements that suggest a continuum—a variation in the degree to which males in a society believe that heterosexual sex is harmful to their health. Enga and Nupe males apparently think that heterosexual sex is harmful to them. Shilluk males think that heterosexual sex would cause harm to their cattle and seemingly to cows’ milk. The Shilluk statements are not clearly about harm to men’s own health. But if we ask “Do people believe that male heterosexuality (even with legitimate partners) brings some harm or danger?” we would have to say that all three of the cultures mentioned had such a belief.

The important point here is that similarities cannot be seen or recognized until we think in terms of *variables*, qualities or quantities that vary along a specified dimension. There is no right or wrong conceptualization of variables; researchers may choose to focus on any aspect of variation. But once researchers perceive and specify similarity, they can perceive and recognize difference. Measurement—deciding how one case differs from another in terms of some scale—is but a short conceptual step away.

Consider now the following ethnographic statements:

4. For the Cuna of Panama, “the sexual impulse is regarded as needing relief, particularly for males, and as an expression of one’s *niga*, a supernatural attribute manifested in potency and strength. On the other hand it is considered debilitating to have sexual relations too often, for this will weaken one’s *niga*” (Stout 1947, 39).
5. And in regard to the Bedouin of Kuwait, “It [sexual intercourse] is the one great pleasure common to rich and poor alike, and the one moment of forgetfulness in his daily round of troubles and hardships that Badawin [Bedouin] or townsmen can enjoy. Men and women equally love the act, which is said to keep a man young, “just like riding a mare” (Dickson 1951, 162).

The Bedouin beliefs contrast most sharply with the beliefs in the other cultures, because heterosexual intercourse appears to be viewed by them as purely pleasurable, with no negative associations. The Cuna seem to be somewhere in the middle. While they view sex as important, they appear to believe that too much is not good.

The variable “degree of men’s fear of sex with women” can be conceptualized as a continuum with gradations.

In a cross-cultural study of this variable, the first author identified four scale points: Societies with mostly negative statements (in the ethnography) about heterosexuality were considered high on men’s fear of sex with women; societies with more or less an equal number of negative and positive statements were considered ambivalent; those with *mostly* positive statements were considered relatively low on men’s fear of sex with women; and those with *only* positive statements were considered as lacking men’s fear of sex with women. While the variable as operationally defined does not capture everything in a culture’s beliefs about heterosexuality, it does capture some distinguishable similarities and differences across cultures (C. R. Ember 1978a).

These examples show that if we focus on a specified aspect or dimension of the variation, similarities and differences become apparent. Framing meaningful and answerable questions, and identifying useful dimensions to do so, is an art. To be meaningful, a question should have some theoretical (generally explanatory) importance. To be answerable, the question should be phrased in a way that allows us to get to an answer on the basis of research. In cross-cultural research, as in most social science research, framing the question (often in a single sentence) is half the battle. Once we frame a question in answerable terms, it is not hard to decide how to go about seeking an answer. And we do not have to limit ourselves to one possible answer; often cross-cultural research involves testing several, not necessarily alternative, answers to the question at issue.

FRAMING THE RESEARCH QUESTION: THE KINDS OF QUESTIONS ASKED

There are at least four kinds of questions we can ask in cross-cultural research:

1. Descriptive/statistical questions. These deal with the prevalence or frequency of a trait. How common is the belief that sex is dangerous to one’s health? What proportion of societies have it? How common is polygyny in the world’s societies?
2. Questions about causes of a trait or custom. Examples are: Why do some societies have the belief that heterosexual sex is harmful? Why do some societies insist on monogamous marriage, whereas most allow polygyny (multiple wives)? Why is war very frequent in some societies and less frequent in others?
3. Questions about the consequences or effects of a particular trait or custom. This kind of question may be phrased broadly: What are the effects of growing up in a society with a great deal of war? Or a consequence question may be phrased much more specifically: What is the effect of polygyny on fertility?
4. Questions that are nondirective and relational. Rather than theorizing about causes or consequences, a researcher may simply ask if a particular aspect of culture is associated with some other aspects. Is there a relationship between type of marriage and level of fertility? Is more war associated with more socialization for aggression in children? No causal direction is specified beforehand with a nondirective relational question.

Of the four types of questions, the descriptive/statistical type is the easiest to address because it tells the researcher what to count in a representative sample of societies. To estimate the frequency of monogamy versus polygyny, we need to establish what each society allows and make a count of each kind of society. The consequence and

relational questions usually specify a set of concrete things to look at. If you want to know whether type of marriage has an effect on or is related to fertility, then you know you need to measure both variables (type of marriage and fertility).

Open-ended causal (and consequential) questions are the most challenging. The only thing specified in the in the open-ended causal question is the dependent variable (the variable to be explained); the only thing specified in the open-ended consequential question is the independent variable (the variable that may have effects). Exactly which variables may be causes or effects is something the investigator has to decide on, often as suggested by some theory. Like a detective who theorizes about suspects and their motives and opportunities, the pursuit of causes involves testing alternative explanations or theories that specify why something is the way it is or how it came to be that way. In science, theories in the literature are usually starting points, but new theories can also be tested.

The basic assumption of cross-cultural research is that comparison is possible because repeated patterns can be identified. Cross-culturalists believe that all generalizations about culture require testing, no matter how plausible we may think they are. This applies to descriptive generalizations presumed to be true (e.g., the presumption that hunter-gatherers are typically peaceful) as well as to presumed relationships or associations (e.g., the presumption that hunting is more likely to be associated with patrilocality). As it turns out, *neither* presumption is generally true of hunter-gatherers (C. R. Ember 1975, 1978b). It is necessary to test all presumed generalizations or relationships because they may be wrong, and we are entitled (even obliged) to be skeptical about any one that has not been tested and supported by an appropriate statistical test.

Cross-culturalists do not believe that cross-cultural research is the only way to test theory about cultural variation. However, such research is viewed as one of the important ways to test theory, if only because cross-cultural research (if it involves a worldwide sample of cases) provides the most generalizable results of any kind of social scientific research. Most cross-culturalists believe in the multi-method approach to testing theory; that is, they believe that worldwide cross-cultural tests should be supplemented (when possible and not impossibly costly) by studying variation in one or more particular field sites (comparisons of individuals, households, communities) as well as within particular regions (e.g., North America, Southeast Asia); theories may also be testable using historical data, experiments, and computer simulations.

Before we discuss the advantages and disadvantages of the various kinds of comparative and cross-cultural research, here is a little historical background.

HISTORY OF CROSS-CULTURAL RESEARCH

The first cross-cultural study was published in 1889 by Edward B. Tylor. In that study, Tylor attempted to relate marital residence and the reckoning of kinship to other customs, such as joking and avoidance relationships. But perhaps because of Francis Galton's objection to Tylor's presentation—see the "Discussion" section at the end of Tylor's paper—little cross-cultural research was done for the next 40 years. (We'll discuss what has come to be called "Galton's Problem" later.) Cross-cultural research started to become more popular in the 1930s and 1940s, at the Institute of Human Re-

lations at Yale. The person who led this rebirth was anthropologist George Peter Murdock (1897–1985). Murdock had obtained his Ph.D. at Yale in a combined sociology/anthropology department that called itself the “Science of Society” (its comparative perspective was established by its founder, William Graham Sumner).

Perhaps the major boost to cross-cultural studies was the Yale group’s development of an organized collection of ethnographic information (first called the “Cross-Cultural Survey,” the precursor of the Human Relations Area Files) that scholars could use to compare the cultures of the world. In the early part of the twentieth century, Sumner had compiled voluminous materials on peoples throughout the world, but his compilation was limited to subjects in which he was personally interested. Later, at the Institute of Human Relations, the group of social and behavioral scientists led by Murdock (including psychologists, physiologists, sociologists, and anthropologists) set out to improve on Sumner’s work by developing the Cross-Cultural Survey. The aim was to foster comparative research on humans in all their variety so that explanations of human behavior would not be culture bound.

The first step was to develop a classification system that would organize the descriptive information on different cultures. This category system became the *Outline of Cultural Materials* (Murdock et al. 2008). The next step was to select well-described cases and to index the ethnography on them, paragraph by paragraph, sometimes even sentence by sentence, for the subject matters covered on a page. As the aim was to file information by subject category to facilitate comparison, and since a page usually contained more than one type of information, the extracts of ethnographic information were typed using carbon paper to make the number of copies needed (corresponding to the number of topics covered); carbon paper was used because the Cross-Cultural Survey antedated photocopying.

In 1946, the Cross-Cultural Survey directorship was turned over to Clellan S. Ford, who undertook to transform it into a consortium of universities. In 1949, first five then eight universities joined together to sponsor a not-for-profit consortium called the Human Relations Area Files, Incorporated (HRAF), with headquarters in New Haven, Connecticut. Over the years, the HRAF consortium added member institutions, and more and more cultures were added to the HRAF collection. In the early days, member institutions received the annual installments of information on xeroxed sheets of paper. Later, the preferred media became microfiche. Since 1994, the only media have been electronic, first CD-ROM and now online (titled *eHRAF World Cultures* [<http://ehrafworldcultures.yale.edu>]). Technological changes have allowed the full-text HRAF Collection of Ethnography to become more and more accessible and more and more efficiently searchable. New cases are added each year. Today, when old cases are added to the electronic HRAF, they are updated if possible. (For more information on the evolution of HRAF, see M. Ember 1997.)

The accessibility that HRAF provides to indexed ethnographic information has undoubtedly increased the number of cross-cultural studies. Considering just worldwide comparisons (the most common type of cross-cultural study), in the years between 1889 and 1947 there were only 10 worldwide cross-cultural studies. In the following 20, years there were 127. And in the next 20 years (ending in 1987), there were 440 (C. R. Ember and Levinson 1991, 138). At present, there are probably 1,000 worldwide

cross-cultural studies in the published literature. (HRAF is compiling a bibliography of cross-cultural studies; this number is just an estimate.)

TYPES OF CROSS-CULTURAL COMPARISON

Cross-cultural comparisons vary along four dimensions: (1) geographical scope of the comparison—whether the sample is worldwide or is limited to a geographic area (e.g., a region such as North America); (2) size of the sample—two-case comparisons, small-scale comparisons (fewer than 10 cases), and larger comparisons; (3) whether the data used are primary (collected by the investigator in various field sites explicitly for the comparison) or secondary (collected by others and found by the investigator in ethnographies, censuses, and histories); and (4) whether the data on a given case pertain to (or date from) just one time period (a *synchronic comparison* of cases) or two or more time periods (a *diachronic comparison*). Although all combinations of the four dimensions are technically possible, some combinations are quite rare. Worldwide cross-cultural comparisons using secondary synchronic data (one “ethnographic present” for each case) are the most common in anthropology.

Comparative research is not just done in anthropology. Worldwide studies using ethnographic data are increasingly done by evolutionary biologists, sociologists, political scientists, and others. Cross-cultural psychologists often compare people in different cultures. And various kinds of social scientists compare across nations. The cross-national comparison is narrower than the worldwide cross-cultural comparison because the results of a cross-national comparison are generalizable only to a limited range of cross-cultural variation—that which encompasses only the complex societies (usually multicultural nation-states) of recent times. The results of a cross-cultural study are generalizable to all types of society, from hunter-gatherers—with populations in the hundreds or a few thousand—to agrarian state societies with populations in the millions—to modern nation-states—with populations in the hundreds of millions.¹

Cross-national research differs from cross-cultural research in ways other than generalizability. Economists, sociologists, and political scientists usually use secondary data when they study samples of nations, but the data are not generally ethnographic. That is, the measures used are not based on cultural information collected by anthropologists or other investigators in the field. Rather, the data used in cross-national comparisons are generally based on censuses and other nationally collected statistics (crime rates, gross national product, etc.), often documented over time. Cross-cultural psychologists are most likely to collect their own (primary) data, but their comparisons tend to be the most limited; very often only two or a few cultures are compared.

Cross-historical studies are still comparatively rare in anthropology (but see Naroll et al. 1974; Peregrine 2001; Peregrine et al. 2004). Few worldwide or even within-region cross-cultural studies have employed data on a given case for more than one time period. But, as noted, some cross-national studies have been cross-historical studies as well, and cross-historical studies using archaeological data are becoming increasingly common (see, e.g., Peregrine 2003; the papers in M. E. Smith 2012). Both cross-national and archaeologically based cross-historical studies can be undertaken by scholars in a wide variety of disciplines (e.g., sociology, history, political science) because there are accessible historical databases. Since primary data are so hard and expensive to collect, it

is hardly surprising that primary comparisons are likely to be small in scale. If you have to collect your data by yourself, going to several places to do so takes a lot of time and money. Large-scale comparisons, which almost always rely on secondary data (collected or assembled previously by others), are generally much less expensive.

Let us turn now to the advantages and disadvantages of the different types of cross-cultural comparison. Our discussion compares the different types with particular regard to theory formulation and theory testing. (Cross-cultural research may also be done to establish the incidence or relative frequency of something.)

Advantages and Disadvantages of the Different Types of Comparison

Worldwide cross-cultural comparisons have two major advantages, compared with other types of comparative research (M. Ember 1991). The major one, as already noted, is that the statistical conclusions drawn from a worldwide comparison of all types of society are probably applicable to the entire ethnographic record, assuming that the sample is more or less free of bias. (See the section on Sampling, below.) This contrasts with results of a within-region comparison, which may or may not be applicable to other regions. And it contrasts with results of a cross-national comparison, which may or may not be applicable to the ethnographic record. The worldwide type of cross-cultural comparison, then, has a better chance than other types of coming close to the goal of knowing that a finding or an observed relationship has nearly universal validity, consistent with the general scientific goal of more and more comprehensive explanations. (Most cross-cultural studies undersample modern industrial societies, but this deficiency will decrease as the ethnographic record increasingly includes the results of ethnographic field studies in industrial countries.)

The other advantage of worldwide cross-cultural comparison is that it maximizes the amount or range of variation in the variables investigated. This may make the difference between a useful and a useless study. Without variation, it's impossible to see a relationship between variables. Even if there's some variation, it may be at just one end of the spectrum of variation. We may think the relationship is positive or negative, because that is all we can observe in one region or in one type of society, but the relationship may be curvilinear in the world, as John Whiting (1954, 524–25) noted. This is what occurs when we plot socioeconomic inequality against level of economic development. There is little socioeconomic inequality in hunter-gatherer societies; there is a lot of inequality in premodern agrarian societies; and inequality is lower again—but hardly absent—in modern industrial societies (M. Ember et al. 1997).

If we only looked at industrial societies, it would appear that more equality goes with higher levels of economic development. But if we only looked just at preindustrial societies, it would appear that more equality goes with *lower* levels of economic development. Thus, to be sure about the nature or shape of a relationship, we have to conduct a worldwide cross-cultural comparison because that shows the maximum range of variation and is the most reliable way to discover the existence and nature of relationships between or among variables pertaining to humans.

When a researcher compares many societies from different parts of the world, she or he is unlikely to know much about each one. If a tested explanation turns out to be supported, the lack of detailed knowledge about the sample cases isn't much of a problem.

However, if the cross-cultural test is disconfirming, it may be difficult to come up with an alternative explanation without knowing more about the particular cases. More familiarity with the cases may help in formulating a revised or new theory that could be tested and supported (Johnson 1991).

Narrowing the scope of a comparative study to a single region may mean that you can know more about the cases and therefore may be more likely to come up with a revised theory if your first tests are unsuccessful. Restricting the study to a region doesn't allow you to discover that the results of the comparison apply to the whole world. (See Burton and White [1991] for more discussion of regional comparisons.) Even a regional comparativist may not know all the cases in the region; that depends mostly on the size of the region. If the region is as large as North America, the comparativist is likely to know less about the cases than if the region studied is the American Southwest. And if you need to look at a sample of the rest of the world to discover how generalizable your results are, you might as well do a worldwide study in the first place!

The objective of large-scale within-region comparisons (using data on all or most of the societies in the region) is usually different from the objective of a worldwide cross-cultural study (Burton and White 1991). Using large numbers of cross-cultural traits, within-region comparativists generally try to arrive at classifications of cultures in order to make inferences about processes of diffusion and historical ancestry. Instead of trying to see how culture traits may be causally related to each other, within-region comparativists are usually more interested in trying to see how the cultures in the region are related to each another. But some regional comparativists are interested in pursuing both objectives at the same time (Jorgensen 1974). The importance of looking at worldwide as well as regional samples, especially if you are interested in relationships between or among variables, is indicated by the following discussion.

Consider the discrepancy between the findings of Driver and Massey (1957) and the findings of M. Ember and C. R. Ember (1971), and Divale (1974) with regard to the relationship between division of labor by gender and where couples live after they get married. Driver and Massey found support in aboriginal North America for Murdock's (1949, 203ff.) idea that division of labor by gender determined matrilineal versus patrilineal residence, but the Embers (and Divale) found no support for this idea in worldwide samples. The Embers found that the relationship varies from region to region. In North America, there is a significant relationship, but in other regions the relationship is not significant. And in Oceania, there is a trend in the opposite direction—matrilineal societies there are more likely to have men doing more in primary subsistence activities. What might account for the difference in the direction of the relationship in different regions? C. Ember (1975) found that the relationship between a male-dominated division of labor and patrilineal residence did hold for hunter-gatherers; hence, she suggested that the frequent occurrence of hunter-gatherers in North America (see Witkowski N.d.) may account for the statistically significant relationship between division of labor and residence in North America.

Fred Eggan (1954) advocated small-scale regional comparisons, which he called "controlled comparisons" because he thought they would make it easier to control on similarity in history, geography, and language. He presumed that the researcher could readily discern what accounts for some aspect of cultural variation within the region if

history, geography, and language were held constant. However, the similarity of cases within a region may be a major drawback.

A single region may not show sufficient variability in the aspect of culture (or presumed causes) the researcher is investigating. Unless a substantial number of the cases lack what you are trying to explain, it would be difficult or impossible to discern what the phenomenon at issue may be related to. For example, suppose almost all the cases in a region share beliefs about sexuality being somewhat harmful. It would be difficult or nearly impossible to be able to figure out what this belief is related to because you could not tell which of the other regularly occurring practices or beliefs in the region might explain the sexual beliefs. Only if some cases lack what you are trying to explain might you see that the hypothetical causes are also generally absent when the presumed effect is absent. Unless there is sufficient variation in all possibly relevant variables, the controlled comparison strategy is a poor choice for testing theory.

Obviously, the controlled comparison is also a poor choice for describing the worldwide incidence of something (unless the region focused on is the only one of interest). While the strategy of controlled comparison may seem analogous to controls in psychology and sociology (which hold some possible causes, and their effects, constant), the resemblance is only superficial. Psychologists and sociologists typically eliminate certain kinds of variation (e.g., in race, religion, ethnicity, or gender) only when they have prior empirical reasons to think that these factors partially predict the variable they are trying to explain. In contrast, those who do controlled comparisons in the anthropological sense usually only *presume* that common history, language, or geography have made a difference. If researchers aren't really controlling on the important predictors when they do a controlled comparison, they aren't necessarily getting any closer to the causal reality by restricting their study to a particular region.

The researcher must collect primary data, in the field if he or she is interested in topics that are rarely (if ever) covered adequately by ethnographers. This was the major reason why John and Beatrice Whiting (see, e.g., B. B. Whiting 1963), who were interested in children's behavior and what it was related to, decided that they had to collect new data in the field for the comparative study that came to be known as the six cultures project. Many aspects of socialization (such as particular practices of the mother) weren't typically described in ethnographies. Similarly, researchers interested in internal psychological states (such as sex-identity, self-esteem, and happiness) couldn't find out about them from ethnographies and therefore would need to collect the data themselves, in the field. How time is allocated to various activities is a nonpsychological example of information that is also not generally covered in ethnographies, or not in sufficient detail.

Although it may always seem preferable to collect primary data as opposed to secondary data, the logistics of cross-cultural comparisons using primary data are formidable in time and expense. And the task of maintaining comparability of measures across sites is difficult (R. L. Munroe and R. H. Munroe 1991a). If a researcher thinks that something like the needed information is already available in ethnographies, a comparison using secondary data is more economical than comparative fieldwork in two or more places. But comparative fieldwork may be the only viable choice when the information needed is not otherwise available.

Similarly, although it may seem preferable to use historical diachronic data to test the temporal ordering implied in causal theories, such data are not often readily available. Because most societies that cultural anthropologists studied lacked native writing, there are usually no historical documents to use for measuring variables for an earlier time. The alternative is to reconstruct the situation in a prior time period using oral history and the occasional documents left by travelers, traders, and other visitors. Such reconstructions are notoriously subject to bias (because of wishful thinking by the researcher). It is also difficult to find diachronic data because different ethnographers have different substantive interests, so different ethnographers who may have worked in the same place at different times may not have collected information on the same variables.

For these reasons, most cross-culturalists think it is more efficient to test causal theories with synchronic data first. If a theory has merit, the presumed causes and effects should generally be associated synchronically. If they are, then we might try to make a diachronic or cross-historical test. If they are, then we might try to see if the presumed causes antedated the presumed effects unless we see first that the synchronic results show correlation.

Diachronic studies may get a lift from *eHRAF Archaeology* (<http://ehrafarchaeology.yale.edu>)—a database with archaeological traditions across the span of prehistory. In recent years, a large body of diachronic archaeological data have been assembled that could be used to test causal theories (Peregrine 2003). Several such tests have been undertaken (e.g. Peregrine 2006; Peregrine et al. 2007). However, linking the archaeological record to cultural traits of interest to cross-culturalists is difficult and often controversial. There are established archaeological correlates of matrilineal descent, sedentarism, and warfare frequency (see Peregrine 2004 for a review), but few others, and some that have been proposed continue to be questioned (e.g., Longacre 1970). In addition, many societies in the HRAF Collection of Ethnography have more than one time focus.² Diachronic data should become increasingly available as the ethnographic record expands with updating and archaeological data are made easier to use, so we should see more cross-historical studies in the future.

SAMPLING

Whatever questions cross-cultural researchers want to answer, they always have to decide what cases to compare. How many cases should be selected for comparison and how should they be selected? If we want to answer a question about incidence or frequency in the world or in a region, it is critical that all the cases (countries or cultures) be listed in the sampling frame (the list to be sampled from, sometimes also called the “universe” or “population”). And all must have an equal chance to be chosen (the major reason sample results can be generalized to the larger universe of cases). Otherwise, it is hard to argue that sample results are generalizable to anything.

It is often not necessary to investigate all cases or even to sample a high proportion of cases from the sampling frame. The necessity of sampling a high proportion of cases depends largely on the size of the population to generalize to. When the population is small, the percentage of cases that must be included to ensure generalizability can be high. When the population is large, the sample size, proportionately, can be quite

small. Except when relationships between contiguous cases are of interest, investigating all the cases would be a colossal waste of time and resources. Political opinion polling is a case in point. The number of voters in the United States is a very large number. Yet, very accurate results can usually be obtained by randomly sampling a few hundred to a few thousand individuals. The size of the sample is not as important as selecting the cases in some random way from a more or less complete list of voters. If you are looking for a large difference between one kind of voter and other, or if the relationship you are examining is strong, you do not have to have a very large sample to obtain statistically significant results. The most important consideration is to sample in an unbiased way, preferably using some kind of random sampling procedure (M. Ember and Otterbein 1991).

Sampling in Comparisons Using Primary Data

There have been relatively few comparisons using primary data (collected by the researcher in two or more field sites) in anthropology. There have been a considerable number of two-case comparisons in cross-cultural psychology, usually comparing subjects in the United States with subjects in some other place. Generally, sampling in comparisons using primary data has been purposive rather than random. That is, the cases compared have not been randomly selected from some sampling frame. This is understandable, given the political realities of gaining permission to do fieldwork in certain countries. In terms of the cost of fieldwork, it is not surprising that two-case comparisons are more common than any other kind of comparison using primary data. Unfortunately, the scientific value of two-case comparisons is dubious. Years ago, Donald Campbell (1961, 344) pointed out that a difference between two cases could be explained by *any* other difference(s) between the cases. Let us consider a hypothetical example.

Assume we are comparing two societies with different levels of fertility. We may think that the difference is due to a need for child labor because much agricultural and household work has to be done. As plausible as this theory may sound, we should be skeptical about it because many other differences between the two societies could be responsible for the difference in fertility. The high-fertility society may also have earlier weaning, a shorter postpartum sex taboo, better medical care, and so on. There is no way, using aggregate or cultural data on the two societies, to rule out the possibility that any of the other differences (and still others not considered here) may be responsible for the difference in fertility. If, however, you have data on a sample of mothers for each society, and measures of fertility and the possible causes for each mother, we could do statistical analyses that would allow us to narrow down the causal possibilities in these two societies. But as suggestive as these results might be, we cannot be sure about what accounts for the difference at issue because we still have only two sample societies.

What is the minimum number of societies for a comparative test using primary data? If two variables are related, the minimum number of cases that might provide a statistically significant result—assuming unbiased sampling, errorless measurement, and a hypothesis that is true—is four (see R. L. Munroe and R. H. Munroe 1991b). Examples of four-case comparisons using primary data, which employed theoretical criteria for case-selection, are the four-culture project on culture and ecology in East

Africa, directed by Walter Goldschmidt (1965), and the Munroes' four-culture project on socialization (R. H. Munroe et al. 1984; R. L. Munroe and R. H. Munroe 1992). In the East Africa project, which was concerned with the effect of ecology/economy on personality and social life, two communities (one pastoral and one agricultural) in each of four cultures (two Kalenjin speaking and two Bantu speaking) were selected. The Munroes selected four cultures from around the world to examine the effects of variation in degree of father-absence and the degree of male-centered social structure.

Sampling in Comparisons Using Secondary Data

You have to decide what your sampling frame is and what list of cases you want to generalize the sample results to. Will it be worldwide (all countries or all societies)? Will it be regional (a broad region like North America, or a narrower one like the cultures of the American Southwest)? When you specify your sampling frame, you're also specifying your unit of analysis.

A country isn't necessarily equivalent to a society or culture in the anthropological sense. A country (or nation-state) is a politically unified population; it may, and often does, contain more than one culture or society. Conventionally, a culture is the set of customary beliefs and practices characteristic of a society, which, in turn, is a population that occupies a particular territory and speaks a common language not generally understood by neighboring populations. Once you know what you want to generalize the sample results to, you should sample from a list containing all the eligible cases. Cross-national researchers have no problem constructing a list of countries. Cross-cultural researchers don't yet have a complete list of the world's described cultures. But there are large lists of cultures to sample from.

Several published lists of societies have served as sampling frames for most cross-cultural studies of the ethnographic record (largely the recent past). A few claim to accurately represent the world's cultures, but we argue below that these claims are problematic and that cross-cultural researchers cannot yet generalize sample results to all cultures. Any claim about a relationship or about the proportion of societies that have a particular trait should be tempered by the recognition that the generalization is only applicable to the list sampled from, and only if the particular cases investigated constitute an unbiased sample of the larger list.

Currently, available cross-cultural samples of the ethnographic record include the following (from largest to smallest): (1) the "Ethnographic Atlas" (1962ff., beginning in *Ethnology* 1 [1962], 113ff. and continuing intermittently over succeeding years and issues of the journal), with a total of 1,264 cases; (2) the "Summary" version of the "Ethnographic Atlas" (Murdock 1967), with a total of 862 cases; (3) the "World Ethnographic Sample" (Murdock 1957), with 565 cases; (4) the *Atlas of World Cultures* (Murdock 1981), with 563 cases; (5) the annually growing HRAF Collection of Ethnography, which covered 385 cultures as of 2013 (the HRAF sample is a collection of texts grouped by culture and indexed by topic for quick information retrieval; no precoded data are provided for the sample cases, in contrast to the situation for all of the other samples except the next one); (6) the "Standard Ethnographic Sample" (Naroll and Sipes 1973 and addenda in Naroll and Zucker 1974), with 273 cases; (7) *eHRAF World Cultures*, which covered 280 cases as of 2013; (8) the "Standard Cross-Cultural Sample"

(Murdock and White 1969), with 186 cases; and (8) the “HRAF Probability Sample” (HRAF 1967; Lagacé 1979; Naroll 1967—also now included and updated in *eHRAF World Cultures*), with 60 cases (this sample is also called the “HRAF Quality Control Sample,” for which some precoded data are available). Before we examine some of the claims made about these various samples, we first need to realize why it is necessary to use random sampling procedures.

According to sampling theory, only random sampling provides an unbiased or representative sample of some larger population or sampling frame (Cochran 1977, 8–11; see also Kish 1987, 16). For example, simple random sampling (using a table of random numbers or a lottery type of selection procedure) guarantees that every case in the sampling frame has had an equal chance to be chosen. (Equiprobability of selection is assumed in most tests that estimate the statistical significance, or likely truth-value, of sample results.) To sample in a simple random fashion, all you have to do is make sure that all cases in the sampling frame are numbered uniquely (no repeats, no cases omitted). Researchers may sometimes choose other kinds of random sampling, such as systematic sampling (every n th case is chosen after a random start) or stratified random sampling (first dividing the sample into subgroups or strata and then randomly sampling from each).

There are two kinds of stratified random sampling. In proportionate stratified random sampling, each subgroup is represented in proportion to its occurrence in the total population; in disproportionate stratified random sampling, some subgroups are overrepresented and others are underrepresented. Disproportionate stratified random sampling is used in cross-cultural research when the researcher needs to overrepresent a rare type of case in order to have enough such cases to study (as in a comparison of relatively rare hunter-gatherers with more common agriculturalists) or when a researcher wants to derive an accurate estimate of some parameter (e.g., mean, variance, or strength of an association) for a rare subgroup. Proportionate random sampling may reduce the sample size needed when there are marked differences between subgroups. However, stratified random sampling may not improve much on the accuracy obtainable with a simple random sample (Kish 1987, 33).

In addition to the samples of the ethnographic record, there are now two samples describing “traditions” in prehistory. The first is a more-or-less complete description of the archaeological traditions in the prehistoric record. All the traditions are described in the *Encyclopedia of Prehistory* (Peregrine and M. Ember 2001–2002) with brief summaries and brief bibliographies. The second is *eHRAF Archaeology*. Modeled after *eHRAF World Cultures*, *eHRAF Archaeology* has documents subject-indexed by paragraph, *eHRAF Archaeology* now contains a random sample of the world’s prehistoric traditions, plus over six complete temporal sequences.

Comparing the Available Samples

Three of the existing cross-cultural ethnographic samples were said to be relatively complete lists at the times they were published. The largest is the complete “Ethnographic Atlas” (with 1,264 cases), published from 1962 on in the journal *Ethnology*. But as its compiler (Murdock 1967, 109) noted, not even the atlas is an exhaustive list of what he called the “adequately described” cultures; he acknowledged that East

Eurasia, the Insular Pacific, and Europe were not well represented.³ For the smaller summary version of the atlas (Murdock 1967), he dropped all cases he considered poorly described. So, if you want your sampling frame to include only well-described cases (in Murdock's opinion), then the 1967 Atlas Summary (with 862 cases) is a reasonable list to sample from.

Raoul Naroll set out to construct a list of societies that met his stringent criteria for eligibility. Some of his criteria were: The culture had to lack a native written language; it had to have an ethnographer who lived for at least a year in the field; and the ethnographer had to know the native language. The resultant sample, which he called the "Standard Ethnographic Sample" (Naroll and Sipes 1973; see also Naroll and Zucker 1974), contains 285 societies; Naroll and Sipes claimed that this list was about 80–90% complete for the cultures that qualified at the time (eastern Europe and the Far East were admittedly underrepresented).

Three of the existing samples (from largest to smallest: the Atlas of World Cultures, the Standard Cross-Cultural Sample, and the HRAF Probability Sample Files) were developed to give equal weight to each of a number of culture areas (areas of similar cultures) in the world. Technically, the samples mentioned in this paragraph are all disproportionate stratified samples (only the HRAF Probability Sample Files uses random sampling to select cases for each culture area identified). The sampling is disproportionate from the strata because the number of cases selected for each identified culture area is not proportionate to the real number of cases in the culture area. The presumption behind all of these stratified samples is that the cultures in a given area are bound to be very similar by reason of common ancestry or extensive diffusion. The designers of these samples wanted to minimize Galton's Problem. In the next-to-last section of this chapter, we discuss whether or not Galton's Problem really is a problem as well as nonsampling solutions to the presumed problem.

There are difficulties with these disproportionate stratified samples. First, exactly how we should define and separate culture areas requires empirical testing; Burton et al. (1996) have shown that social structural variables do not cluster consistently with Murdock's major cultural regions. Second, disproportionate stratified sampling is a less efficient way to sample (i.e., it requires more cases) than simple random sampling. Third, even if the disproportionate sample uses random sampling from each stratum, every case selected will not have had an equal chance to be chosen. This makes it difficult or impossible to estimate the commonness or uniqueness of a particular trait in the world. If we do not know how common a trait is in each culture area, we cannot correct our counts by relative weighting, which we would need to do to make an accurate estimate of the frequency of the trait in the world.

Many have used all or some of the cases in the Standard Cross-Cultural Sample (SCCS) (Murdock and White 1969) for cross-cultural studies, at least partly because the published literature contains a large number of codes (ratings of variables) on those cases; many of these codes were reprinted in Barry and Schlegel (1980). This sample is claimed to be representative of the world's known and well-described cultures (as of 1969), but that claim is dubious for two reasons. First, disproportionate sampling does not give an equal chance for each culture to be chosen. Second, the single sample case from each cluster was chosen judgmentally, not randomly. However, Gray (1996) com-

pared results from the SCCS with a thousand random samples from the Ethnographic Sample and found little evidence of bias in the SCCS. He claims that the only bias found was the inclusion of better-described societies. (Judgmental criteria were also used to choose the five cases per culture area for the *Atlas of World Cultures* [Murdock 1981].)

Of the three ethnographic samples discussed here, the HRAF Probability Sample is the only one employing *random* sampling within strata (the 60-culture sample includes a random selection from each identified culture area). However, the other criteria for selection were so stringent (e.g., at least 1,200 pages of cultural data focused on a community or other delimited unit; substantial contributions from at least two different authors) that only 206 societies in the whole world were eligible for inclusion as of the time the sample was constructed (in the late 1960s).

Two other samples should be briefly discussed, because researchers have used them as sampling frames in cross-cultural studies. One is the World Ethnographic Sample (Murdock 1957). In addition to being based on judgmental sampling of cases within culture areas, it has one other major drawback. A time and place focus is not specified for the cases, as in the two Ethnographic Atlas samples (full and summary), the Standard Cross-Cultural Sample, and the Standard Ethnographic Sample. If researchers want to use some of the precoded data for the World Ethnographic Sample, they would have no idea what the “ethnographic present” is for a case. As we discuss later, focusing on the same time and place for all the measures on a case is usually called for in testing for an association; you may be introducing error if you do not make sure that the measures used pertain to the same time and place for the case.

Finally, let’s turn to the entire HRAF Collection of Ethnography. Like most of the other samples, it, too, was based on judgmental selection. But because it covers many cultures all over the world and provides ethnographic materials that are complexly indexed for rapid information retrieval, the collection has often been used as a sampling frame for cross-cultural studies. (As of 2013, the HRAF collection covered 385 cultures, at least 44% of the world’s well-described cultures if you go by Murdock’s [1967] total of 862 in the summary version of the Ethnographic Atlas.)

The major advantage of the HRAF collection, as compared with the other available lists or samples of cultures, is that only HRAF provides ethnographic texts on the cases. The other samples provide only coded data (usually) and only limited bibliography (usually). If the codes constructed by others don’t directly measure what you are interested in, but you use them anyway, you may be reducing your chances of finding relationships and differences that truly exist. Hence, if you need to code new variables or you need to code something in a more direct way, you are likely to do better if you yourself code from the original ethnography (C. R. Ember et al. 1991). Library research to do so would be very time consuming, which is why HRAF was invented in the first place.

If you use the HRAF collection, you don’t have to devote weeks to constructing bibliographies for each sample case; you don’t have to chase down the books and other materials you need to look at, which might otherwise have to be obtained by interlibrary loan; and you don’t have to search through every page of a source (that often lacks an index) to find all the locations of the information you seek. The HRAF collection gives you the information you want on a particular topic, from all of the sources

processed for the culture, in a single place. That place, with the electronic HRAF, is now online. If you want to examine the original ethnography on a case (with all the context), and particularly if you want to construct your own measures, there is no substitute for the HRAF collection.

If you are starting out to do your first cross-cultural study, how should you sample? If you want to use some of the data already coded for one of the available samples, by all means use that sample *as your sampling frame*, particularly if it is one of the larger lists (such as the summary version of the Ethnographic Atlas [Murdock 1967]). The sampling frame becomes the list you are claiming to generalize to. (A larger claim that you are generalizing to the world is inappropriate.)

If you want to code all of your variables yourself, you can do so most economically by sampling from the HRAF collection. If a sample size of 60 is large enough, the HRAF Probability Sample has the advantage of giving you a randomly selected case from each of the 60 culture areas. If you want to code some variables yourself and use some pre-coded variables, you can sample from the intersection between HRAF and the sample with the pre-coded variables. Whatever sampling frame you use, you should select your cases in some standard random fashion, because only random sampling entitles you to infer that your statistically significant sample results are probably true for the larger universe. If, for some reason, you cannot sample randomly from some list, be sure to avoid selecting the cases yourself. After a random sample, the next best thing is a sample constructed by others (who could not know the hypotheses you want to test). The wonderful thing about a random sample is that wherever your research stops, after 20 or 40 or 200 randomly selected cases, you will always be entitled to conclude that a statistically significant result in your sample is probably true for the larger universe.

MEASUREMENT

How you choose to measure some variable of interest to you depends at least partly on your implicit or explicit theory. After all, why are you interested in variable X in the first place? Your implicit or explicit theory specifies which variables are of interest and provides a model of how they may be related. Theories are generally evaluated by testing hypotheses derived from them. (What does the theory imply in the way of relationships?) The variables in a test can be fairly specific, such as whether or not a culture has a ceremony for naming a newborn child, or they may be quite abstract, such as whether the community is harmonious. Whether the concept is fairly specific or not, no variable is ever measured directly. We are so used to a thermometer measuring heat that we may forget that heat is an abstract 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 don't see heat; we see only the movement of the substance in the confined space. So all measurement is indirect. But some measures are better (more direct, more predictive, fewer errors) than others.

The three most important principles in designing a measure are: (1) try to be as specific as possible in deciding how to measure the theoretical variable you have in mind; (2) try to measure the variable as directly as possible; and (3) if possible, try to measure the variable in a number of different ways.

The first principle recognizes that science depends on replication; if we are to be confident about our findings, other researchers must be able to repeat them. To facilitate replication, the original researchers have to be quite explicit about what they intended to measure and exactly how they measured it.

The second principle recognizes that although all measurement is indirect, some measures are more direct than others. To measure how “rainy” an area is, you could count the number of days that it rains while you are there for a week’s vacation and then multiply by 52. But it would be better to count the number of rainy days, on average, over a number of years, and it would be best if you measured the *total number of inches* of rain per years, on average, over a number of years.

The third principle of good measurement is that few measures exactly measure what they are supposed to measure, so it is better to use more than one way to measure the theoretical variable of interest. However, much of cross-cultural research is limited by what is available in ethnographies. So, the availability of relevant information is the major constraint on the number of supposedly equivalent measures one might construct.

Measures have to be specified for each variable in the hypothesis. Devising a measure involves at least four steps: (1) theoretically defining the variable of interest (in words or mathematically); (2) operationally defining the variable, which means spelling out the empirical information needed to make a decision about where the case falls on the scale that the researcher has devised for measuring it; (3) pretesting the measure to see if it can be applied generally (to many if not most cases) (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 too often lacking, the measure needs to be rethought); and (4) performing reliability and validity checks (reliability involves the consistency, replicability, and stability of a measure; validity involves the degree to which the measure reflects what it is supposed to reflect; for more discussion of these issues; see C. R. Ember et al. 1991). Because most attention has been paid to measurement issues in secondary comparisons, we focus mainly on them in what follows.

To illustrate processes involved in measurement, let’s consider that a researcher has an idea about why many societies typically have extended families. Although the concept of extended families may appear straightforward, it needs to be defined explicitly. The researcher needs to decide whether to focus on extended family households or to include extended families that are not co-residential. The choice should depend on the theory. If the theory discusses labor requirements that would favor an extended family staying together (see Pasternak et al. 1997, 237–39), then extended family households should be measured.

When this is decided, the researcher still needs to state what an “extended family” means and what a “household” means. And she or he has to decide on the degree (relative frequency) to which a sample society has extended family households. The first thing to decide is what is meant by an extended family. The researcher may choose to define a family as a social and economic unit consisting minimally of at least one or more parents and children; an extended family might be defined as consisting of two or more constituent families united by a blood tie; and an extended family household might be defined as an extended family living co-residentially—in one house,

neighboring apartments, or a separate compound. Having defined the concepts, the researcher must then specify the counting procedure—how to measure the degree to which a society has extended family households. 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. For example, suppose by degree (of extended “family-ness”) we operationally mean the percentage of households in the community that are extended families. The range of possible scale scores is from 0 to 100%. Suppose further that we instruct our coders to rate a case only if the ethnographer specifies a percentage or we can calculate a percentage from a census of the households. If we did a pretest, we would find that very few ethnographers tell us the percentage of extended family households or 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.” Our operational definition of percentage of extended family households, although perfectly worthy, may not be that useful if we cannot find enough societies with reports based on household censuses.

What can we do? There are three choices. We can stick to our insistence on the best measure and study only societies for which a percentage is given (or can be calculated); we may have to expand our search (enlarge our sample) to find enough cases that have such precise information. Or we can redesign our measure to incorporate descriptions in words that are not based on census materials. Or we can choose not to do the study because we can’t measure the concept exactly how we want to.

Faced with these choices, most cross-cultural researchers would opt to redesign the measure so as to incorporate word descriptions. Word descriptions do convey information about degree, even if not so precisely. If an ethnographer says “extended family households are typical,” we don’t know if that means 50% or 100%, but we can be very confident it does not mean 0–40%. And we can be fairly sure it does not mean 40–49%. If the relative frequency of extended families (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. A measure, going by words, might read something like what follows.

Code extended family households 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.
Don't know if there is no information on form of household, or there is contradictory information.

The next step is to pretest this measure, preferably with coders who haven't had anything to do with creating the scale. Four distinctions may be too difficult to apply to the word descriptions usually found in ethnographies, so a researcher might want to collapse the scale a little. Or, two coders may not agree with each other frequently. If so, the investigator may have to spell out the rules a little more. And if we decide to use the scale described above, what do we do when the ethnography actually gives us 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). For instance, if 70% of the households have extended families, and 30% are independent, we would choose scale score 3. But we might decide to use two scales: a precise one based on numerical measurement (percentages) for cases with numbers or percentages, the other scale relying on words (when the ethnography provides only words). C. R. Ember et al. (1991) recommend using 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 one. (The less-precise scale should be less accurate than the more precise one, assuming that the former sometimes has to rely on ambiguous words.) Stronger results with the more precise scale would increase our confidence that the relationship observed even with the less-precise one is true. We discuss the issue of validity in the next section and the issue of reliability in the section titled *Minimizing Coder Error*.

How to Minimize Error in the Design of Measures

Two kinds of measurement errors are usually distinguished: systematic and random errors. They have different effects on data analysis and each type is handled differently (Zeller and Carmines 1980, 12).

Systematic error or bias exists if there is a consistent, predictable departure from the "true" score. Examples would be a scale that inflates everyone's weight by half a pound or a tendency by observers to ignore certain types of aggression in behavior observations. In cross-cultural research, systematic error can come from ethnographers, informants, design errors in measurement, or coders. Ethnographers may not mention something or may underreport it (e.g., not mention the number of Western objects in the village). Or they may overreport (e.g., overemphasize the unilineality of the descent system to fit reality into an ideal type). Informants may over- or underreport (e.g., not mention an illegal or disapproved activity). Coders may interpret ethnography from the point of view of their own culture, their gender, or their personality. As discussed below, we may be able to detect possible bias by hypothesizing and testing for it. However, one major

type of systematic error can't be detected so easily—the error that's introduced because the measure consistently under- or overestimates the theoretical variable (Blalock 1968; Cook and Campbell 1979, 64; Zeller and Carmines 1980, 11).

Random error, which is error in any direction, weakens correlations (Blalock 1972, 414). Naroll (1962) called random errors “benign,” perhaps because he and other social scientists commonly worry more about accepting a relationship as true that is false (Type I error) than about failing to accept a true relationship (Type II error). As Naroll (1962) pointed out, systematic error in two variables in the same direction could conceivably create a result when none is really there. (C. R. Ember has unpublished results showing that systematic error has to be enormous to produce a significant result when there really is none.) However, if our purpose is to find relationships when they are really there, we should take steps to minimize random error. But let us turn first to the problem of systematic error that is due to the lack of fit between the theoretical construct and the measure that presumes to tap that construct.

Even though all measurement is indirect so validity can't ever be established beyond a doubt (Campbell 1988), some measures are more direct and therefore more likely to be valid than others. More direct measures involve little question that they are measuring what they are supposed to measure. 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 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 based on explicit ethnographers' reports of work patterns rather than to infer the work pattern from general statements about how husbands and wives get along.

Some measures pose few validity problems. Where the operationalization is very close to the theoretical variable, for example, the sex of a person (Blalock 1968, 20), the measure requires so little inference that there is little question about its validity. Other measures seem to have “face validity” too—few researchers would question their validity. For example, if the theoretical variable is the rule of residence in a society, we usually think that the ethnographer's identification of the rule of residence is an operational measure with high face validity, despite the apparent lack of agreement between ethnographers Ward Goodenough (1956) and John Fischer (1958). The disagreement between them over Chuuk (Truk) appears to have been interpreted by some anthropologists as an indication that a fieldworker's conclusions are bound to be subjective and therefore unreliable.

We think that there are two reasons why this interpretation is incorrect. First, the two investigators were in the field at different times, and practices can vary even over just a few years. Second, both Goodenough and Fischer found the Chuukese to be predominantly matrilocal (Goodenough 71%, Fischer 58%). Indeed, they differed only with respect to 13% of the cases, which Goodenough classified as avunculocal and Fischer as patrilocal. (In some people's eyes, avunculocal residence is patrilocal because living with husband's mother's brother is living with *husband's* relatives.) With respect to the major or predominant residence pattern, the cross-culturalist would not misclassify the Chuukese using Fischer's *or* Goodenough's data, even though they apparently disagreed on some details.

Unlike experimental psychologists, cross-cultural researchers using secondary data are not able to use a great variety of validation techniques. There is rarely a standard measure to evaluate a new measure against. (If there were, we would always use the standard as the measure.) If there is little information in ethnographies on particular topics, it is sometimes difficult to think of multiple measures of the same theoretical variable. Given this situation, it is best to use measures that are the most direct and therefore have the highest face validity.

However, many concepts of interest to cross-cultural researchers are not easily measured directly. Two situations in which indirect measures might justifiably be used are when the theoretical variable can be measured only “projectively” (e.g., unconscious fear of something as measured by how often, proportionately, folktales mention it) or when very few ethnographies give information allowing a more direct measure. Unconscious variables, in particular, do not lend themselves to direct measurement. Psychologists (and others) employ projective testing and disguised measures when subjects cannot or will not give honest answers to certain questions. For example, a field investigator (comparative or not) cannot simply ask a boy “Do you wish to be a woman?” and expect to get an answer that will reflect the boy’s degree of feminine identification, any more than a survey researcher can ask “Are you prejudiced against blacks?” and get an accurate response most of the time.

Cross-cultural researchers may conclude, therefore, that culturally shared personality dispositions might be more accurately coded from folktales than from presumptions by ethnographers about unconscious feelings. Although a researcher may well be aware that a more direct measure would be preferable, more indirect measures may also be chosen because a more direct measure may be usable only for a small proportion of cases. Whatever the reason, the decision to employ an indirect or proxy measure should be made only if the investigator can justify its use on the basis of explicit and strong reasoning about why we should accept the validity of the proxy. More on this point below.

When the cross-culturalist decides to use a more indirect measure, we strongly recommend that he or she develop direct measures for some proportion of the cases (even if only a minority) to evaluate the validity of the more indirect measure. If it’s not possible to correlate the two measures to validate the more indirect or proxy measure (because both kinds of information are hardly ever available for a particular case), the researcher could still see if the more indirectly measured cases show weaker correlations than the more directly measured cases. As noted above, if the more direct measure produces higher coefficients of association than the more indirect measure, and the directions of the results are similar with both measures, we can be more confident that the indirect measure is probably tapping the same variable as the more direct one.

Researchers sometimes choose proxy measures over more direct measures because the proxy measures are readily available in precoded databases. But unless it is shown that they correlate with more direct measures (which requires going back to the original ethnographies), we’re skeptical about the validity of opportunistic proxy measures. We acknowledge that there’s enormous value in having databases with codes provided by previous researchers. It isn’t always necessary to code things anew, and using available codes when they fit your interests can allow time to code additional variables. But

the practice of using precoded variables as proxies, without any attempts to validate them, deserves our suspicion. (Of course, the investigators originally responsible for the available codes are not to be blamed for how others use them.)

The designer of a measure also needs to consider the degree to which the information required is available in ethnographies. No matter how direct a measure may seem conceptually, it may require a high degree of inference by the coder if there is little relevant information in ethnographies. The terms “high-inference” and “low-inference” variables, first introduced by J. W. M. Whiting (1981), are useful for discussing this aspect of measurement design. Variables that require low inference on the part of the coder tend to deal with visible traits or customs, usually reported by ethnographers, and are easily located in ethnographies (Bradley 1987, 1989; Burton and White 1987; J. W. M. Whiting 1981; and see White 1990).

High-inference variables often require complex coding judgments and are therefore difficult to code reliably. The codings of low-inference variables are less likely to contain error, and independent coders are therefore more likely to agree on codings. For example, Bradley (1987) compared her codings of presence versus absence of the plow with Pryor’s (1985) codings for 23 societies; there was 96% agreement between the two data sets. The only disagreement was about a case that Pryor himself expressed uncertainty about. Thus, presence or absence of the plow, which ethnographers can observe and record without interpretation, is a low-inference variable. Others include the type of carrying device for infants, shapes of houses, domestic animals and major crops, and many elements of material culture (Bradley 1987; Burton and White 1987; J. W. M. Whiting 1981). Note that the dimension of low versus high inference may be uncorrelated with the dimension of more versus less direct measurement. Presence or absence of the plow may be measurable with low inference but if you consider it to be a proxy measure for the judgment that men do most of the agricultural work, then your measure would be low-inference but indirect.

Because the measurement of some variables requires moderate levels of inference, such measures are more subject to random error. Relevant information is also missing more often, therefore you are likely to be able to code only a small proportion of the sample. In addition, when variables require moderate levels of inference, coders usually agree with each other less. Bradley (1989) presents evidence that coding the gender division of labor in agriculture requires moderate inference. Other examples of variables that require a moderate degree of inference are the proportion of day an infant is carried (J. W. M. Whiting 1981) and warfare frequency (C. R. Ember and M. Ember 1992a; Ross 1983). For moderate-inference variables, coders usually have to read through a considerable amount of ethnographic material that is not explicitly quantitative in order to rate a case. Because of the imprecision, the coding decision is more likely to contain some error.

The highest degree of inference is required when researchers are interested in assessing general attitudes or global concepts such as “evaluation of children” (Barry et al. 1977). Such global concepts don’t have obvious empirical referents to guide coders, and it’s easy to see how coders focusing on different domains might justifiably code the same society differently. The most appropriate solution here, we believe, is to develop a series of more specific measures with clear empirical referents, as Whyte (1978) did

for the various meanings of women's status and Ross (1983) did for the various dimensions of political decision making and conflict.

Random errors may be more likely if the investigator does not precisely specify time and place for all variables. Divale (1975), acting on the Embers' suggestion, has shown with a few examples how lack of time and place focus, which presumably increases random error, tends to lower correlations. The same time and place for a sample case should be attended to whether previous codes are used or new codes are developed. For example, M. Ember (1974, 1984/85) has presented evidence that polygyny is favored by a shortage of men because of high male mortality in war. If polygyny were present in a particular society as of 1900, but you measured the sex ratio in a later ethnography (after warfare ceased), you would be likely to find a more or less balanced sex ratio at the later time.

Would this case be exceptional to the theory that high male mortality (and an excess of women) favors polygyny? The answer is yes, but it would not be appropriate to measure the two variables in this way. Each of the two variables should be measured synchronically (for more or less the same time period), or you could measure male mortality in war or the sex ratio for a slightly earlier time than you measure form of marriage. Otherwise, you may be introducing so much measurement error that the relationship between excess women and polygyny could be masked (C. R. Ember and M. Ember 2009, 76). (C. R. Ember et al.'s [1992] concordance between cross-cultural samples can help you match time and place foci across samples.) Requiring that your measurements on each case pertain to the same time and place (or the appropriate relative times for a diachronic [over time] test) can only maximize your chances of seeing a relationship that truly exists.

Minimizing the Effect of Ethnographer (or Informant) Error

Measurement error caused by ethnographer or informant error is not often considered something we can deal with in the measurement process. But that is not necessarily true, Naroll (1962) proposed methods to deal with such errors, and others have developed additional methods.

The supposedly poor or uneven quality of the ethnographic record is often cited as invalidating cross-cultural research. This notion is puzzling, given the usual high regard ethnographers have for their own work. If most anthropologists have high regard for their own work, how could the bulk of ethnography be poor unless most anthropologists were deluding themselves (C. R. Ember 1986, 2)? Certainly there are errors in the ethnographic record, and we must try to minimize their effects, but the critics' worry may derive from their ignorance about the effect of error on results.

Space here does not allow it, but we could show statistically that even a great deal of random error hardly ever produces a statistically significant finding. And even systematic error would not normally produce a statistically significant finding that was false. (There is always the possibility of deliberate cheating; but this probably does not happen often and other investigators' attempts to replicate a result will eventually reveal it.) Statistically speaking, random error generally reduces the magnitude of obtained correlations. This means that more error lessens the likelihood of finding patterns that are there. And if error makes us *less* likely to infer statistical significance, it should not

be assumed that significant cross-cultural results are generally invalid. It may seem paradoxical, but the more random error there is, the more likely the “true” results are better than the observed results.

Naroll (1962) proposed an indirect method—data quality control—to deal with systematic informant and ethnographer errors. His basic procedure involved identifying factors that might produce biases in the reporting of certain variables (e.g., a short stay in the field would presumably make for underreporting secret practices such as witchcraft). Indeed, Naroll found that short-staying ethnographers were significantly less likely to report witchcraft than were long-staying ethnographers. This suggests the possibility of systematic error due to length of stay. However, as Naroll himself was aware, there are other possible interpretations of the correlation. One is that short stays may be more likely in more complex cultures. If more complex cultures are less likely to have witchcraft beliefs than less complex ones, the correlation between length of stay and the presence of witchcraft would be spurious, not due to systematic underreporting by short-staying ethnographers.

Still, data quality factors could account for some results. For example, to exclude the possibility that a data quality factor may account for a correlation because that factor is correlated with both variables in the correlation, Naroll advised the researcher to control statistically for the data quality factor. Naroll’s concern with the possible influence of data quality persuaded some researchers to test for systematic biases of various kinds (gender of ethnographer, type of training, knowing the native language, etc.) in evaluating their results. Naroll believed that the coding of information on qualities of the ethnographer and on conditions of the fieldwork should be a regular part of the coding process.

But C. R. Ember et al. (1991) do not agree, for two reasons. First, as Naroll (1977) also noted, it is very expensive for researchers to code for a large number of features that could, but probably do not, produce false correlations. Second, of the large number of studies done by Naroll, his students, and others (see Levinson [1978] for substantive studies employing data quality controls), hardly any have found that a data quality feature accounts for a correlation [but see also Divale 1976; Rohner et al. 1973]). Therefore, Ember et al. (1991) recommend that before investigating relationships between data quality variables and substantive variables, researchers should have plausible theoretical reasons for thinking they may be related. If we cannot imagine how a data quality variable could explain a correlation, it is not necessary to spend time and money coding for it. Only plausible alternative predictors should be built into a research design. However, Ember et al. (1991) recommend that researchers develop a specific data quality code for each substantive variable (for each case), to provide a direct assessment of the quality of the data in regard to that variable.

To illustrate the suggested procedure, compare the information in the following three statements, which a coder might use to measure the frequency of polygyny among married men: “Polygyny is the form of marriage that men aspire to,” “Only the senior men have more than one wife,” and “The household survey indicates that 15% of the married men are married polygynously.” Although none of these statements may contain error, and we can’t assume that the one based on quantitative information is the most accurate, a coder trying to rate frequency of polygyny according to an ordinal

scale would have the most trouble coding the first statement because it tells us only that polygyny is present and preferred (at least by men).

The researcher could try to minimize error by instructing the coders not to make frequency judgments based on statements about what people prefer (ideal culture). But the researcher could also develop a data quality code for the measure of frequency of polygyny. The highest-quality score would be given to a code based on a census. The lowest-quality score would be given to information pertaining to ideal culture (such as the first statement listed above) or to a judgment based on inference (e.g., polygyny is inferred to be not so common because only senior men are said to have more than one wife). A middle-quality score might be given to information such as in the following statement: “The typical married man has only one wife.”

With a data quality code by case for each variable in a correlational test, we could analyze the results with *and* without the “poorer” quality data. The omission of cases with scores based on poor-quality data (e.g., vague ethnographic statements) should yield stronger results than the data set that includes poor-quality data. (For how cases with poorer data quality can be omitted, see our discussion toward the end of the next section on Minimizing Coder Error.) And because the standard errors will be higher with more random error, our chances of finding a statistically significant relationship are greater with higher-quality data. We think the data quality control procedure suggested here provides two important advantages. First, it taps the quality of the ethnographic information more directly than Naroll’s suggested strategy, which may not make any difference at all in regard to a particular correlation. Second, the data quality coding we suggest can be done quite efficiently, at the same time substantive variables are coded, because reading additional material is not required.

Another problem the ethnographic record poses to cross-cultural researchers is that in addition to errors in what is reported, there may also be problems about what is not reported. Ethnographers may not have gone to the field with a comprehensive guide to what kinds of information could be collected, such as is provided by the *Outline of Cultural Materials* (Murdock et al. 2008). For this and other reasons, ethnographers often pay little or no attention to a question that is of interest later to the cross-cultural researcher. What should a cross-culturalist do?

We don’t recommend inferring that something is absent if it is not reported, unless the cross-cultural researcher can be quite sure that it would have been reported if it had been present. For example, if an ethnographer didn’t mention puberty rites but thoroughly discussed childhood and adolescence, absence of puberty rites could reasonably be inferred. If, however, the ethnographer didn’t collect any information on adolescence, the fact that no puberty rites are mentioned shouldn’t be taken to mean that they were absent. Researchers need to specify coding rules for inferring absence (see the measure of extended family households described above) or they need to instruct their coders not to make inferences.

A further strategy to deal with missing data is to interview the original ethnographers themselves (or others who have worked for extended periods in the society) to supplement the information not present in the published sources (Levinson 1989; Pryor 1977; Ross 1983). The cross-cultural researcher has to be careful to keep to the

same time and place foci of the published data; if you call a recent ethnographer about a case, you should ask only about the time and place foci of the published information.

Finally, if data on some of the sample societies are missing, the cross-cultural researcher may decide to impute missing values. Burton (1996) and Dow and Eff (2009) have described and evaluated a number of procedures for doing so. The cross-cultural researcher needs to remember that any method of imputation is likely to increase measurement error. Therefore, the advantage of imputation (to increase sample size) has to be weighed carefully against the possible increase of error. Researchers who impute some data should consider doing analyses with and without the imputed data to see if the imputing has misleadingly improved the results, just as we can compare data sets with and without dubious codings to see if including them has transformed a borderline or nonsignificant result into a significant one.

Minimizing Coder Error

The coding process itself can produce measurement error. If the investigator is also the coder, there is the possibility of systematic bias in favor of the theory being tested (Rosenthal 1966; Rosenthal and Jacobson 1968). For that reason alone, many researchers prefer to use “naive” or theory-blind coders. However, naive coders may not be as likely as experienced coders to make accurate judgments. Experienced researchers have skills that should make for more accurate coding because they are more likely to be aware that an ethnographer’s words should not always be taken at face value. For example, an experienced researcher is more likely to know that avunculocal residence might be called patrilocal residence, and that hunter-gatherers may get plenty of food (even if they have to move their camps frequently, it doesn’t necessarily mean their food supply is precarious). Furthermore, experienced coders are more likely to pay attention to time and place foci—to know that when one ethnographer describes what Samoans do, he or she may not be talking about the particular time period or particular group of Samoans a previous ethnographer has studied and described (M. Ember 1985).

Bradley (1987) argues that naive coders can make systematic errors when coding instructions are insufficiently precise, especially when coding high-inference variables. For example, differences between her codes for the division of labor in agriculture and those of the coders for Murdock and Provost (1973) could be explained by the possibility that the Murdock and Provost coders were not instructed to consider differences between crop types or which phase of the agricultural sequence was to be coded. Naive coders might also be more likely to make judgments that are systematically biased toward their own cultural assumptions, as suggested by the experimental findings presented by D’Andrade (1974).

Researchers can try to minimize the error of inexperienced coders by being as explicit as possible in their coding instructions and by making sure that the codes do not surpass the information that is generally available in the ethnographic literature (Tatje 1970). Coding should have to make as few inferential leaps as possible. The process of trying to spell out all the possible obstacles to coding is an important part of the research design. It may be that having at least one relatively inexperienced coder provides an advantage—it may force the researcher to be as clear as possible in the operationalization of theoretical concepts.

Researchers may worry that coders will introduce systematic errors because of their gender, their political ideology, their personality, or their faulty assumptions about different types of societies. But such contamination may not be so likely. Whyte (1978) did not find more significant relationships between gender of coder and his many indicators of women's status than would be expected by chance. His research suggests that systematic coding bias is most likely to occur when codes are very general (requiring a high degree of inference), which may allow the coder's personal background to exert an influence on the coding process. Whyte suggests that personal/cultural biases can be avoided if coders are asked to rate concrete or specific customs and behaviors.

Designing studies to test systematically for coder biases is normally quite expensive because to do so properly requires more than one coder for each bias-type. Thus, it's more cost effective for the investigator and a naive coder to rate the cases. Not only could we then compare the two sets of ratings for reliability, we could also see if both sets of ratings give similar results. If they do not, that would be something to worry about. It might only be that the naive coder's ratings contained more error; the results using only that coder's ratings should be weaker than the results using only the investigator's ratings. Perhaps the best strategy is to test hypotheses using only those cases that both raters agreed on. That way, you would probably be omitting the cases with more random error.

This brings us to the concept of interrater reliability—the extent to which different persons using the same measure achieve the same score or the same relative ranking for each case rated (Nunnally 1967, 172; Zeller and Carmines 1980, 6). Researchers who assess interrater reliability usually show a correlation coefficient between the two raters' judgments for at least a sample of the rated cases. Or, they may show the percentage of agreement. Both of these measures have advantages and disadvantages. The percentage of agreement may detect systematic error, whereas a correlation coefficient may not. Suppose one coder always gives a score that's one point higher score than the other. The correlation coefficient will be perfect (1.00); the percentage of agreement will be zero.

On the other hand, the percentage of agreement measure can't distinguish between substantial disagreements and small disagreements. On a 10-point scale, a disagreement of 1 point counts the same as a disagreement of 10 points (Rohner and Katz 1970, 1068). But percentage of agreement will detect differences between coders due to systematic inflation or deflation by one of the coders, whereas this kind of systematic bias will not affect a reliability correlation. So, it is useful to compute both kinds of measures of interrater reliability.

There is no clear decision point as to what is an acceptably high coefficient or percentage of agreement. Coefficients and percentages of agreement over .80 appear to be considered good and are usually reported without comment; more than .70 appears to be considered minimally acceptable; less than .70 leaves a feeling of unease. Neither method of measuring interrater reliability is a good way to deal with the situation where one rater does not assign a scale score (says "don't know") and the other rater does. Correlation coefficients are easily computed; if some form of Pearson's r is used, it is easy to interpret—the square of the coefficient equals the proportion of variance explained. We always expect that there will be some interrater inconsistency (a study without any at all would probably be suspect). How do researchers deal with disagreements?

If a researcher does a reliability check on a small portion of the cases and the reliability coefficient is reasonably high, she or he usually uses the scores of the one coder who rated all of the cases. If a researcher has two or more coders for each case, then a variety of strategies may be followed. Scores may be summed or averaged, or disagreements may be resolved by discussion. An advantage of the summing or averaging method is that both coders are given equal voice; personality differences between the coders cannot influence resolutions. A second advantage is that the “effective reliability” is increased. Rosenthal and Rosnow (1984, 163–65) indicate that if the correlation coefficient between two judges is .75, the reliability of the mean of the two judges’ ratings is actually higher (.86). This is a better estimate of the reliability of the measure; it is an example of the Spearman-Brown prophecy formula, also known as Cronbach’s alpha (Nunnally 1978, 211–16; Romney 1989).

A disadvantage of the summing or averaging procedure is that one coder may be a better coder (more careful, more knowledgeable about the material), and the rating by that person might objectively deserve more weight. If coders are asked to discuss their disagreements and come to a resolution, the more knowledgeable coder might point to information that the other coder missed. This strategy might be particularly useful when each coder has read a large amount of ethnographic material.

There are also some disadvantages in the resolution method. First, as mentioned, one coder may have undue influence over the other. Certainly, if one coder was the investigator and the other a paid assistant, the resolution method might not be unbiased. But even if the coders were roughly equal in status, one personality may dominate the other. Rohner and Rohner (1981) discuss a procedure for testing for the influence of one coder over another. However, a major problem with their method is that it cannot distinguish between influence because of power and influence because of more information. Another disadvantage is that it may add measurement error if the coders feel obliged to come to some resolution, even when the data are too ambiguous to justify resolution.

The Embers have found that results get stronger when they use only those cases with the most reliable initial scores.⁴ This should not be surprising. The most reliable scores are presumably those that independent coders will agree on initially, before any attempts to resolve disagreements. When we eliminate cases with disagreement, our results should get stronger because we are probably eliminating the more ambiguous cases, which are probably the ones more likely to be coded inaccurately. An interrater reliability coefficient may be acceptable, but you could still have a lot of cases that have been measured inaccurately. Even though the size of the sample is reduced when we eliminate the least reliable scores in some way, our chances of finding a true relationship are improved.

We think the best way to maximize the reliability of ratings and results is to use only those ratings (and cases) that the independent coders initially rated in exactly the same way (or very similarly), before any attempts at resolution. The researcher could provide a code that tells the reader how closely the raters agreed initially on the variable for a particular case (see C. R. Ember and M. Ember 1992b). This kind of reliability code (by variable, by case) would allow subsequent users of the data code to choose their own degree of reliability. Just as it is likely that results are more robust when we omit cases

that did not have higher quality data, so they should be more robust when we omit cases that were not rated in much the same way by the coders initially. Chances are, more ambiguous ethnographic information will generally be coded with more error, and it does us no good to cloud the situation by including them in our analyses.

Minimizing Error Due to Sampling: Galton's Problem

Another major reason some question cross-cultural findings is referred to as "Galton's Problem." In 1889, Francis Galton heard Edward Tylor's presentation of what is generally considered the first cross-cultural study. Galton (see Tylor 1889, 270–72) suggested that many of Tylor's cases were duplicates of one another because they had similar histories, so Tylor's conclusions were suspect because the sample size was unjustifiably inflated. Since the 1960s, Raoul Naroll and others such as James Schaefer, Colin Loftin, Malcolm Dow, and E. Anthon Eff have considered Galton's Problem a serious threat to cross-cultural research. They have devised several methods to test for the possible effects of diffusion and historical relatedness (Naroll 1970; for earlier references, see C. R. Ember 1990; more recent references can be found in Dow 2007, 2008; Dow and Eff 2008, 2009; Eff 2004). The concern behind these methods is that statistical associations may not be valid if the correlations could be attributed mostly to diffusion (cultural borrowing) or common ancestry.

How serious is Galton's Problem? Cross-culturalists disagree (see M. Ember and Otterbein 1991 for references; see also C. R. Ember 1990). Most but not all cross-culturalists think Galton's Problem is a serious one (see the names in the previous paragraph). We and others (Strauss and Orans, Otterbein—for references, see C. R. Ember 1990) think it is not serious, albeit for different reasons. A great deal of the disagreement hinges on different theoretical assumptions about causality—if you believe that cultures are slow to change and people are strongly influenced by their past history and by their neighbors, the more important you think the problem is. If you think that cultures change as circumstances change and borrow because something is perceived to be functional or adaptive, Galton's Problem is not perceived as important.

Statistical tests require two kinds of independence, which we can call sampling independence and measurement independence (Blalock 1972; Kish 1965, 1987). Sampling independence requires that every case has an equal chance to be chosen. Measurement independence requires that the measure for one or more variables for one case is not influenced by the measures of another case. Copying answers on an exam is a clear case of nonindependence. Blalock (1972, 144–45) suggests that crime rates for different census tracts may be nonindependent if crimes are committed by some of the same individuals across boundaries.

It is measurement independence that Galton appeared most concerned with; the assumption is that common history created similarity. However, some who argue against Galton's Problem (M. Ember and Otterbein 1991, 223–24) assert that usually different societies have mutually unintelligible languages and therefore their speech communities have been separated for at least 1,000 years. If two related languages began to diverge 1,000 or more years ago, many other aspects of the cultures will also have diverged. Therefore, such cases could hardly be duplicates of each other. If you push Galton's Problem to the limit and avoid any two cases that share a common

history and language, then psychological studies with more than one individual per culture would be suspect!

Until recently, whether or not you worried about Galton's Problem made a big difference about how you would do a study. Naroll's tests for the possibility of diffusion were quite time consuming to carry out; probably for this reason, most cross-culturalists altered their sampling strategy to eliminate multiple cases from the same culture area. For example, the Standard Cross-Cultural Sample (Murdock and White 1969) and the HRAF Probability Sample both contain only one culture per identified culture area. Unfortunately, this solution is not ideal because sampling from geographical clusters does not give all societies an equal chance to be chosen (a principle of statistical independence). Another sampling strategy is to take a relatively small random sample from a large list, such as the Ethnographic Atlas (Murdock 1962ff.) or the summary Ethnographic Atlas (Murdock 1967). By chance, there are likely to be only a few closely related cases; if there are, some can be eliminated randomly (C. R. Ember and M. Ember 2009, 109; M. Ember and Otterbein 1991).

Recently, however, mathematical anthropologists have developed statistical solutions and computer programs that treat the proximity of societies (in distance or language) as a variable whose influence can be tested in a multiple regression analysis. (This is called "testing for spatial autocorrelation." For newer treatments based on network autocorrelation, see Dow 2007, 2008; Eff 2008; Eff and Dow 2009).

Whether or not a researcher agrees that Galton's Problem is a problem, the recent mathematical and computer solutions do not require a special sampling strategy, nor do they require expensive time-consuming controls. If you worry about Galton's Problem, all you have to do is test statistically for the possibility that proximity or common ancestry accounts for a result (Burton et al. 1996; Dow et al. 1984). Even without a statistical control for autocorrelation, cross-culturalists who randomly sample from a larger sampling frame can redo their analyses by randomly omitting more than a single case from the same culture area. If the results do not change substantially after multiple cases from an area are omitted, the original result cannot be due to duplication of cases. Indeed, duplication may weaken results because a set of historically related cases may be exceptional to a cross-cultural generalization rather than consistent with it. (For some evidence on this possibility, see M. Ember 1971.)

MAXIMIZING THE INFORMATION VALUE OF STATISTICAL ANALYSIS

Most cross-cultural studies aim to test hypotheses, so some kind of inferential statistic is usually used to make a decision about whether the tested hypothesis should be accepted or rejected. This is where statistical researchers resort to the concept of level of significance. Cross-culturalists generally conform to the social science convention of accepting a hypothesis (at least provisionally) if there were five or fewer chances out of a hundred ($p < .05$) of getting the result just by chance. That is, if 100 different random samples were examined, only 5 or fewer would show the same result or one stronger.

Early cross-cultural studies usually relied on contingency tables and chi-square as a test of significance for the relationship between two variables (bivariate analyses), but with the advent of statistical software packages for personal computers, increasing numbers of cross-culturalists realized that they could achieve more powerful and

more informative statistical results in testing hypotheses if they measure ordinally (or intervally) rather than nominally. In addition, multivariate analyses, such as logistic regression (for nominal dependent variables) and multiple regression (for interval and sometimes ordinal dependent variables), can be used to evaluate the independent effects of two or more variables.⁵

Nominal measurement is putting a case into an appropriate set without implying that one set is higher or lower than another on some scale (e.g., female vs. male; extended family households vs. independent family households). Often, nominal variables can be appropriately transformed to ordinal variables. Rather than classify societies as just having independent family households versus extended family households, our discussion earlier showed how to measure the degree to which a society has extended family households in terms of a four-point ordinal scale. In ordinal-scale measurement, a higher (or lower) number implies more (or less) of the variable measured.

Cross-cultural researchers can rarely measure variables in terms of interval or ratio scales, where the distance between two adjacent numbers is equal to the distance between any other two adjacent numbers (the ratio-scale has a meaningful zero-point). But there are plenty of possible interval or ratio scales: number of people in the community, population density, average rainfall, average mean temperature (the Fahrenheit or Celsius scale is an interval scale because zero does not mean the absence of temperature), altitude, and others. Labovitz (1967, 1970) has suggested that statistical tests originally designed for interval-level data may be used with ordinal data when the number of ordered scale scores is not very small. For example, C. R. Ember and M. Ember (1992a, 1994) used multiple regression analysis when the dependent variables had five or more ordinal scale scores. Indeed, cross-cultural researchers are now more likely to use multivariate techniques to discover the relative effects of two or more predictors. These techniques are especially important for evaluating alternative theories that seem to be equally supported by the bivariate results.

Most cross-culturalists don't use mathematical formulas to decide on sample size in advance of their research. This is a pity, for a researcher often spends a good deal of effort trying to code a large number of cases in the belief that a large sample size is necessary. Generally, you don't need a large sample to obtain a significant result. A small one can give you a trustworthy or significant result if it is strong. A large sample is necessary if you want to detect a weak association or effect. Indeed, with a large enough sample, you will often find trivial effects that may hinder interpretation of results. Nominal-level tests (like chi-square) require the largest sample sizes; ordinal and interval-level tests generally require smaller sample sizes.

Kraemer and Tiemann (1987) have provided a master table to calculate approximately what sample size you will need, if you can specify how big an effect or correlation you are looking for, and how much possibility of error you can tolerate. (They also provide specific formulas for particular measures and tests of significance.) Suppose a researcher is looking for a correlation of .50 or better and wants to be 90% confident that he or she has found a true result, and one that has only a .05 chance of being false. The approximate number of cases required would be 30. If the researcher insists on a *p* value of .01, 45 cases would be required. To detect a weak correlation of only .24, a *p* value of .05 requires 144 cases; a *p* value of .01 requires 219 cases (Kraemer and Thiemann 1987).

Statistical techniques (factor analysis, multidimensional scaling, correspondence analysis) can evaluate whether there are one or more dimensions in a set of related measures. These techniques are especially useful for constructing complex scales from multiple indicators (e.g., indicators of social complexity, status of women) and for exploring patterns in the data. Most of these techniques are not used for hypothesis testing.

While cross-culturalists aim to reject causal theories when the hypotheses derived from them are not supported by correlational tests, they generally cannot differentiate between causes and effects even when hypotheses are supported. However, much more can be done than usually is. First, partial correlation and path analysis are designed to help evaluate alternative causal models. Only a few cross-cultural studies have used such techniques (C. R. Ember and M. Ember 1992a, 1994; Kitahara 1981). Second, more cultures than you might think have been studied for more than one time period.⁶ Cross-culturalists can use this diachronic information to study whether changes occur in time as predicted by our causal theories, but there's been little research of this kind as yet. (More will probably be done as the HRAF Collection of Ethnography increasingly includes information on more than one time period in the history of a case.) In the future, cross-archaeological studies will also be used to test theories diachronically (M. Ember and C. R. Ember 1995) as *eHRAF Archaeology* grows and contains more tradition sequences.

Causal theories can also be tested using “phylogenetic” methods (Borgerhoff Mulder et al. 2001; Mace and Holden 2005). Phylogenetically based comparative methods employ an existing phylogeny that describes a hypothetical relationship between cultures based on some known variable, most commonly language. The known variable is used to create a phylogenetic tree upon which specific cultural traits are mapped. The tree of mapped traits is then examined or compared to other mapped traits to determine whether or not a specific hypothesis is supported. Because the trees are based on known relationships, they can be used to infer the evolution of various traits found among a group of cultures; that is, cultures grouped in a phylogeny based on language that also share a particular cultural trait might be hypothesized to have inherited that trait from the common ancestral culture that spoke the ancestral language (e.g., see the papers in Mace et al. 2005). Since the phylogenetic tree also hypothesizes ancestral states, phylogenetic methods can also be used as a method of reconstructing ancient cultural systems (e.g., Currie et al. 2010; Jordan 2011).

Phylogenetically based comparative methods are not without controversy (Moore 1994), and their use to date has been limited both in geographic scope and scholarly impact. However, they have the potential to allow cross-culturalists to explore complex evolutionary and causal relationships. They also are unique among cross-cultural methods in avoiding Galton's Problem altogether, as defining the relationships between cases (i.e., the extent to which they are independent) is a foundation of the method (Borgerhoff Mulder et al. 2001). Because of these strengths, we anticipate phylogenetically based comparative methods will be more widely used in the coming years.

CONCLUSION

Cross-cultural researchers do not deny the uniqueness of particular cultures, but they look at cultures in a different way—focusing on qualities or quantities that vary along some specified dimension. These variables do not capture everything about cultural

attitudes, beliefs, values, or behaviors, but they do exhibit some distinguishable similarities and differences. A large part of the art of cross-cultural research is learning how to focus on dimensions of variation; so is learning how to frame a meaningful and answerable question. Questions range from the descriptive, dealing with the prevalence or frequency of a trait, to questions about causes, consequences, or relationships without specified causal direction.

In addition to believing that comparison is possible, cross-culturalists generally assume that all generalizations require testing on some unbiased sample of cases. There are many types of cross-cultural comparison, from regional to worldwide, small or large, using primary or secondary data, synchronic or diachronic. They all have advantages and disadvantages. Large samples aren't necessarily better than small ones—what is most important is that the studied cases should fairly represent the universe of cases to which the results are generalizable; some kind of random sampling is the best way to ensure representativeness. Because theoretical variables are never measured directly, we have devoted a lot of space to various issues of measurement, including how to minimize error in designing measures and how to minimize the effects of ethnographer and coder error in cross-cultural tests using ethnographic data. We advocate the design of the most direct measures possible, those that require the least coder inference, and data quality scores for each variable for each culture.

Cross-cultural researchers don't unanimously agree about everything they do. They disagree about the seriousness of Galton's Problem, and how to deal with it, for instance. And they disagree often about causal interpretations. But with all their disagreement, they agree that it is necessary to do cross-cultural research (preferably a variety of cross-cultural studies—worldwide, regional, primary, secondary, synchronic, diachronic) to arrive at trustworthy and comprehensive explanations of human behavior, explanations that are probably true because they apply to the vast majority of cultures. We need to test our explanations in other ways too—within cultures, ethnohistorically, cross-historically, experimentally, by computer simulations. But cross-cultural research (of the usual synchronic variety) is a necessary part of the social scientific enterprise because only a cross-cultural test gives us a relatively low-cost opportunity to discover that a theory or explanation does not fit the real world of cultural variation and therefore should be rejected, at least for now.

NOTES

1. Although it might be argued that the Standard Cross-Cultural Sample does not adequately cover contemporary national cultures, it does contain some cultures that constitute the core of countries today—Japanese, Koreans, Russians, Egyptians, Vietnamese, Chinese, and Haitians. Considering that the number of well-described cultures probably exceeds 2,000, we would not expect that many to be national cultures.
2. Swanson (1980) reported that about one-third of the probability sample files cases were well described for more than one time period; since then, more recent time frames have been added. Researchers can use C. R. Ember et al.'s (1992) cross-cultural concordance that matches time and place foci across samples to find cultures described for more than one time period.
3. However, Bondarenko et al. (2005) added cultures of Russia.
4. This was the case in the results reported in C. R. Ember and M. Ember 1992a. Although the initial results that included all resolved ratings were not shown, they were considerably

weaker with all resolved ratings. The reported results were based on omitting the less reliably rated variables.

5. Autocorrelation methods referred to Galton's Problem is a form of multiple regression.
6. See note 2.

REFERENCES

- Barry, H., III, L. Josephson, E. Lauer, and C. Marshall. 1977. Agents and techniques for child training; cross-cultural codes 6. *Ethnology* 16: 191–230.
- Barry, H., III, and A. Schlegel, eds. 1980. *Cross-cultural samples and codes*. Pittsburgh: University of Pittsburgh Press.
- Blalock, H. M., Jr. 1968. Measurement problem: A gap between the languages of theory and research. In *Methodology in social research*, ed. H. M. Blalock and A. B. Blalock, 5–27. New York: McGraw Hill.
- Blalock, H. M., Jr. 1972. *Social statistics*, 2nd ed. New York: McGraw-Hill.
- Bondarenko, D., A. Kazankov, D. Khaltourina, and A. Korotayev. 2005. A corrected ethnographic atlas. *Ethnology* 44: 261–89.
- Borgerhoff Mulder, M., M. George-Cramer, J. Eshleman, and A. Ortolani. 2001. A study of East African kinship and marriage using a phylogenetically based comparative method. *American Anthropologist* 103: 1059–82.
- Bradley, C. 1987. *Women, children and work*. Ph.D. dissertation. Irvine: University of California Press.
- Bradley, C. 1989. Reliability and inference in the cross-cultural coding process. *Journal of Quantitative Anthropology* 1: 353–71.
- Burton, M. L. 1996. Constructing a scale of female contributions to agriculture: Methods for imputing missing data. *Cross-Cultural Research* 30: 3–23.
- Burton, M. L., C. C. Moore, J. W. M. Whiting, and A. K. Romney. 1996. Regions based on social structure. *Current Anthropology* 37: 87–123.
- Burton, M. L., and D. R. White. 1987. Cross-cultural surveys today. *Annual Reviews of Anthropology* 16: 143–60.
- Burton, M. L., and D. R. White. 1991. Regional comparisons, replications, and historical network analysis. *Behavior Science Research* 25: 55–78 (Special issue: Cross-Cultural and Comparative Research: Theory and Method).
- Campbell, D. T. 1961. The mutual methodological relevance of anthropology and psychology. In *Psychological anthropology: Approaches to culture and personality*, ed. F. L. K. Hsu, 333–52. Homewood, IL: Dorsey Press.
- Campbell, D. T. 1988. *Methodology and epistemology for the social sciences: Selected papers*. Chicago: University of Chicago Press.
- Cochran, W. G. 1977. *Sampling techniques*, 3rd ed. New York: John Wiley.
- Cook, T. D., and D. T. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Currie, T. E., S. J. Greenhill, R. D. Gray, T. Hasegawa, and R. Mace. 2010. Rise and fall of political complexity in island South-East Asia and the Pacific. *Nature* 467: 801–4.
- D'Andrade, R. 1974. Memory and the assessment of behavior. In *Measurement in the social sciences*, ed. T. Blalock, 149–86. Chicago: Aldine-Atherton.
- Dickson, H. R. P. 1951. *The Arab of the desert: A glimpse into Badawin life in Kuwait and Sau'di Arabia*. London: George Allen & Unwin.
- Divale, W. T. 1974. Migration, external warfare, and matrilineal residence. *Behavior Science Research* 9: 75–133.

- Divale, W. T. 1975. Temporal focus and random error in cross-cultural hypothesis tests. *Behavior Science Research* 10: 19–36.
- Divale, W. T. 1976. Female status and cultural evolution: A study in ethnographer bias. *Behavior Science Research* 10: 19–36.
- Divale, W. T. 2007. Galton's problem as multiple network autocorrelation effects: Cultural trait transmission and ecological constraint. *Cross-Cultural Research* 41: 336–63.
- Divale, W. T. 2008. Network autocorrelation regression with binary and ordinal dependent variables. *Cross-Cultural Research* 42: 394–419.
- Dow, M. M., M. L. Burton, D. White, and K. Reitz. 1984. Galton's problem as network autocorrelation. *American Ethnologist* 11: 754–70.
- Dow, M. M., and E. A. Eff. 2008. Global, regional, and local network autocorrelation in the standard cross-cultural sample. *Cross-Cultural Research* 42: 148–71.
- Dow, M. M., and E. A. Eff. 2009. Cultural trait transmission and missing data as sources of bias in comparative survey research: Explanations of polygyny re-examined. *Cross-Cultural Research* 43: 134–51.
- Driver, H., and W. C. Massey. 1957. Comparative studies of North American Indians. *Transactions of the American Philosophical Society* 47: 165–456.
- Eff, E. A. 2004. Does Mr. Galton still have a problem? Autocorrelation in the standard cross-cultural sample. *World Cultures* 15: 153–70.
- Eff, E. A. 2008. Weight matrices for cultural proximity: Deriving weights from a language phylogeny. *Structure and Dynamics: eJournal of Anthropological and Related Sciences* 3(2). <http://escholarship.org/uc/item/13v3x5xw> (accessed September 15, 2012).
- Eff, E. A., and M. M. Dow. 2009. How to deal with missing data and Galton's problem in cross-cultural survey research: A primer for R. *Structure and Dynamics: eJournal of Anthropological and Related Sciences* 3(2). <http://www.escholarship.org/uc/item/7cm1f10b> (accessed September 15, 2012).
- Eggan, F. 1954. Social anthropology and the method of controlled comparison. *American Anthropologist* 56: 655–63.
- Ember, C. R. 1975. Residential variation among hunter-gatherers. *Behavior Science Research* 10: 199–227.
- Ember, C. R. 1978a. Men's fear of sex with women: A cross-cultural study. *Sex Roles: A Journal of Research* 4: 657–78.
- Ember, C. R. 1978b. Myths about hunter-gatherers. *Ethnology* 17: 439–48.
- Ember, C. R. 1986. The quality and quantity of data for cross-cultural studies. *Behavior Science Research* 20: 1–16.
- Ember, C. R. 1990. Bibliography of cross-cultural methods. *Behavior Science Research* 24: 141–54.
- Ember, C. R., and M. Ember. 1992a. Resource unpredictability, mistrust, and war: A cross-cultural study. *Journal of Conflict Resolution* 36: 242–62.
- Ember, C. R., and M. Ember. 1992b. Warfare, aggression, and resource problems: Cross-cultural codes. *Behavior Science Research* 26: 169–226.
- Ember, C. R., and M. Ember. 1994. War, socialization, and interpersonal violence: A cross-cultural study. *Journal of Conflict Resolution* 38: 620–46.
- Ember, C. R., and M. Ember. 2009. *Cross-cultural research methods*, 2nd ed. Lanham: AltaMira.
- Ember, C. R., and D. Levinson. 1991. The substantive contributions of worldwide cross-cultural studies using secondary data. (Special issue. Cross-cultural and comparative research: Theory and method.) *Behavior Science Research* 25: 79–140.

- Ember, C. R., with the assistance of H. Page, Jr., T. O'Leary, and M. M. Martin. 1992. *Computerized concordance of cross-cultural samples*. New Haven, CT: Human Relations Area Files. Printed on demand.
- Ember, C. R., M. H. Ross, M. L. Burton, and C. Bradley. 1991. Problems of measurement in cross-cultural research using secondary data. (Special issue. Cross-cultural and comparative research: Theory and method.) *Behavior Science Research* 25: 187–216.
- Ember, M. 1971. An empirical test of Galton's problem. *Ethnology* 10: 98–106.
- Ember, M. 1974. Warfare, sex ratio, and polygyny. *Ethnology* 13: 197–206.
- Ember, M. 1984/85. Alternative predictors of polygyny. *Behavior Science Research* 19: 1–23.
- Ember, M. 1985. Evidence and science in ethnography: Reflections on the Freeman-Mead controversy. *American Anthropologist* 87: 906–10.
- Ember, M. 1991. The logic of comparative research. (Special issue. Cross-cultural and comparative research: Theory and method.) *Behavior Science Research* 25: 143–53.
- Ember, M. 1997. Evolution of the Human Relations Area Files. *Cross-Cultural Research* 31: 3–15.
- Ember, M., and C. R. Ember. 1971. The conditions favoring matrilineal residence versus patrilineal residence. *American Anthropologist* 73: 571–94.
- Ember, M., and C. R. Ember. 1995. Worldwide cross-cultural studies and their relevance for archaeology. *Journal of Archaeological Research* 3: 87–111.
- Ember, M., C. R. Ember, and B. Russett. 1997. Inequality and democracy in the anthropological record. In *Inequality, democracy, and economic development*, ed. M. Midlarsky, 110–32. Cambridge: Cambridge University Press.
- Ember, M., and K. F. Otterbein. 1991. Sampling in cross-cultural research. (Special issue. Cross-cultural and comparative research: Theory and method.) *Behavior Science Research* 25: 217–35.
- Ethnographic Atlas. 1962–. *Ethnology* 1:113ff. and intermittently thereafter.
- Fischer, J. L. 1958. The classification of residence in censuses. *American Anthropologist* 60: 508–17.
- Goldschmidt, W. 1965. Theory and strategy in the study of cultural adaptability. *American Anthropologist* 67: 402–8.
- Goodenough, W. 1956. Residence rules. *Southwestern Journal of Anthropology* 12: 22–37.
- Gray, J. P. 1996. Is the Standard Cross-Cultural Sample biased? A simulation study. *Cross-Cultural Research* 30: 301–15.
- HRAF. 1967. The HRAF quality control sample universe. *Behavior Science Notes* 2: 63–69.
- Johnson, A. 1991. Regional comparative field research. (Special issue. Cross-cultural and comparative research: Theory and method.) *Behavior Science Research* 25: 3–22.
- Jordan, F. 2011. A phylogenetic analysis of the evolution of Austronesian sibling terminologies. *Human Biology* 83: 297–321.
- Jorgensen, J. G. 1974. On continuous area and worldwide sample cross-cultural studies. In *Comparative studies by Harold E. Driver and essays in his honor*, ed. J. G. Jorgensen, 195–204. New Haven, CT: HRAF Press.
- Kish, L. 1965. *Survey sampling*. New York: John Wiley.
- Kish, L. 1987. *Statistical design for research*. New York: Wiley.
- Kitahara, M. 1981. Men's heterosexual fear due to reciprocal inhibition. *Ethos* 9: 37–50.
- Kraemer, H. C., and S. Theimann. 1987. *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.
- Labovitz, S. 1967. Some observations on measurement and statistics. *Social Forces* 46: 151–60.
- Labovitz, S. 1970. The assignment of numbers to rank order categories. *American Sociological Review* 35: 515–24.

- Lagacé, R. O. 1979. The HRAF probability sample: Retrospect and prospect. *Behavior Science Research* 14: 211–29.
- Levinson, D. 1978. Holocultural studies based on the Human Relations Area Files. *Behavior Science Research* 13: 295–302.
- Levinson, D. 1989. *Family violence in cross-cultural perspective*. Newbury Park, CA: Sage.
- Longacre, W. 1970. *Archaeology as anthropology: A case study*. Anthropological papers of the University of Arizona, Number 17. Tucson: University of Arizona Press.
- Mace, R., and C. Holden. 2005. A phylogenetic approach to cultural evolution. *TRENDS in Ecology and Evolution* 20: 116–21.
- Mace, R., C. Holden, and S. Shennan, eds. 2005. *The evolution of cultural diversity: A phylogenetic approach*. London: UCL Press.
- Meggitt, M. J. 1964. Male-female relationships in the highlands of Australian New Guinea. *American Anthropologist* 66: 202–24.
- Moore, J. H. 1994. Putting anthropology back together again: The ethnogenetic critique of cladistic theory. *American Anthropologist* 96: 925–48.
- Munroe, R. H., H. S. Shimmin, and R. L. Munroe. 1984. Gender understanding and sex role preference in four cultures. *Developmental Psychology* 20: 673–82.
- Munroe, R. L., and R. H. Munroe. 1991a. Comparative field studies: Methodological issues and future possibilities. (Special issue. Cross-cultural and comparative research: Theory and method.) *Behavior Science Research* 25: 155–85.
- Munroe, R. L., and R. H. Munroe. 1991b. Results of comparative field studies. (Special issue. Cross-cultural and comparative research: Theory and method.) *Behavior Science Research* 25: 23–54.
- Munroe, R. L., and R. H. Munroe. 1992. Fathers in children's environments: A four culture study. In *Father-child relations*, ed. B. S. Hewlett, 213–29. New York: Aldine de Gruyter.
- Murdock, G. P. 1949. *Social structure*. New York: Macmillan.
- Murdock, G. P. 1957. World ethnographic sample. *American Anthropologist* 59: 664–87.
- Murdock, G. P. 1962ff. Ethnographic atlas. *Ethnology* 1 ff.
- Murdock, G. P. 1967. Ethnographic atlas: A summary. *Ethnology* 6: 109–236.
- Murdock, G. P. 1981. *Atlas of world cultures*. Pittsburgh: University of Pittsburgh Press.
- Murdock, G. P., C. S. Ford, A. E. Hudson, R. Kennedy, L. W. Simmons, and J. W. M. Whiting. 2008. *Outline of cultural materials*, 6th rev. ed. with modifications. New Haven, CT: HRAF Press.
- Murdock, G. P., and C. Provost. 1973. Factors in the division of labor by sex: A cross-cultural analysis. *Ethnology* 12: 203–25.
- Murdock, G. P., and D. R. White. 1969. Standard Cross-Cultural Sample. *Ethnology* 8: 329–69.
- Nadel, S. F. 1954. *Nupe religion*. London: Routledge & Kegan Paul.
- Naroll, R. 1962. *Data quality control—a new research technique: Prolegomena to a cross-cultural study of culture stress*. New York: Free Press.
- Naroll, R. 1967. The proposed HRAF probability sample. *Behavior Science Notes* 2: 70–80.
- Naroll, R. 1970. Galton's problem. In *A handbook of method in cultural anthropology*, ed. R. Naroll and R. Cohen, 974–89. Garden City, NY: Natural History Press.
- Naroll, R. 1977. Cost-effective research versus safer research. *Behavior Science Research* 11: 123–48.
- Naroll, R., V. L. Bullough, and F. Naroll. 1974. *Military deterrence in history: A pilot cross-historical pilot survey*. Albany: State University of New York Press.
- Naroll, R., and R. G. Sipes. 1973. Standard ethnographic sample. *Current Anthropology* 14: 111–40.
- Naroll, R., and H. Zucker. 1974. Reply. *Current Anthropology* 15: 316–17.

- Nunnally, J. C. 1967. *Psychometric theory*. New York: McGraw-Hill.
- Nunnally, J. C. 1978. *Psychometric theory*, 2nd ed. New York: McGraw-Hill.
- Pasternak, B., C. R. Ember, and M. Ember. 1997. *Sex, gender, and kinship: A cross-cultural perspective*. Upper Saddle River, NJ: Prentice Hall.
- Peregrine, P. 2001. Cross-cultural comparative approaches in archaeology. *Annual Review of Anthropology* 30: 1–18.
- Peregrine, P. 2003. Atlas of cultural evolution. *World Cultures* 14: 2–88.
- Peregrine, P. 2004. Cross-cultural approaches in archaeology: Comparative ethnology, comparative archaeology, and archaeoethnology. *Journal of Archaeological Research* 12: 281–309.
- Peregrine, P. 2006. Synchrony in the New World: An example of ethnoarchaeology. *Cross-Cultural Research* 40: 6–17.
- Peregrine, P. 2007. Modeling state origins using cross-cultural data. *Cross-Cultural Research* 41: 1–12.
- Peregrine, P., C. R. Ember, and M. Ember. 2004. Universal patterns in cultural evolution: An empirical analysis using Guttman scaling. *American Anthropologist* 106: 145–49.
- Peregrine, P., and M. Ember, eds. 2001–2002. *Encyclopedia of prehistory*. New York: Kluwer Academic/Plenum Publishers.
- Pryor, F. L. 1977. *The origins of the economy: A comparative study of distribution in primitive and peasant economies*. New York: Academic Press.
- Pryor, F. L. 1985. The invention of the plow. *Comparative Studies in Society and History* 27: 727–43.
- Rohner, R. P., B. R. DeWalt, and R. C. Ness. 1973. Ethnographer bias in cross-cultural research: An empirical study. *Behavior Science Notes* 8: 275–317.
- Rohner, R. P., and L. Katz. 1970. Testing for validity and reliability in cross-cultural research. *American Anthropologist* 72: 1068–73.
- Rohner, R. P., and E. C. Rohner. 1981. Assessing interrater influence in cross-cultural research: A methodological note. *Behavior Science Research* 16: 341–51.
- Romney, A. K. 1989. Quantitative models, science and cumulative knowledge. *Journal of Quantitative Anthropology* 1: 153–223.
- Rosenthal, R. 1966. *Experimenter effects in behavioral research*, enl. ed. New York: Irvington.
- Rosenthal, R., and L. Jacobson. 1968. *Pygmalion in the classroom*. New York: Holt, Rinehart, and Winston.
- Rosenthal, R., and R. L. Rosnow. 1984. *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.
- Ross, M. H. 1983. Political decision making and conflict: Additional cross-cultural codes and scales. *Ethnology* 22: 169–92.
- Seligman, C. G., and B. Z. Seligman. 1932. *Pagan tribes of the Nilotic Sudan*. London: George Routledge & Sons.
- Smith, M. E., ed. 2012. *The comparative archaeology of complex societies*. Cambridge: Cambridge University Press.
- Special Issue. Cross-Cultural and Comparative Research: Theory and Method. 1991. *Behavior Science Research* (now *Cross-Cultural Research*), 25 (1–4).
- Stout, D. B. 1947. *San Blas Cuna acculturation: An introduction*. New York: Viking Fund Publications in Anthropology.
- Swanson, E. C. 1980. A note about temporal foci in the HRAF Probability Sample Files. *Cross-Cultural Research* 15: 295–307.
- Tatje, T. A. 1970. Problems of concept definition for comparative studies. In *A handbook of method in cultural anthropology*, ed. R. Naroll and R. Cohen, 689–96. Garden City, NY: Natural History Press.

- Tylor, E. B. 1889. On a method of investigating the development of institutions applied to the laws of marriage and descent. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 18: 245–72.
- White, D. R. 1990. Reliability in comparative and ethnographic observations: The example of high inference father-child interaction measures. *Journal of Quantitative Anthropology* 2: 109–50.
- Whiting, B. B., ed. 1963. *Six cultures: Studies of child rearing*. New York: Wiley.
- Whiting, J. W. M. 1954. The cross-cultural method. In *Handbook of social psychology*, Vol. 1, ed. G. Lindzey and E. Aronson, 523–31. Cambridge, MA: Addison-Wesley.
- Whiting, J. W. M. 1981. Environmental constraints on infant care practices. In *Handbook of cross-cultural human development*, ed. R. H. Munroe, R. L. Munroe, and B. B. Whiting, 155–79. New York: Garland.
- Whyte, M. K. 1978. Cross-cultural studies of women and the male bias problem. *Behavior Science Research* 13: 65–80.
- Witkowski, S. R. N.d. Environmental familiarity and models of band organization. Unpublished manuscript. New Haven, CT: Human Relations Area Files.
- Zeller, R. A., and E. G. Carmines. 1980. *Measurement in the social sciences: The link between theory and data*. New York: Cambridge University Press.

Copyrighted Material
Not for Reproduction