Abstract:
This article discusses the construction and uses of databases in comparative research.

The World Cultures Database

Douglas R. White
Department of Anthropology, 3151 Social Science Plaza, University of California, Irvine; drwhite@uci.edu

This article discusses the construction and uses of databases in comparative research.

Keywords: Databases, Comparative Research

1. INTRODUCTION

The richness of Anthropology and Sociology lie in Ethnography, Ethnohistory, and conceptualizations to match the unfolding of human behavior. The comparative analyst is especially fortunate in that this database, while constantly subject to revision, is potentially cumulative and holistic. Its cumulative quality comes from the limited universe of well described societies, and either continuous area or standard samples which focus secondary studies time and again on the same units, so that the set of variables constructed for analytic purposes accumulates across research projects dealing with the same social units. The holistic quality comes from the attempt to restore the context that is lost in treating each case (1) in terms of a limited set of analytic variables, and (2) as a separate or independent unit rather than as a series of facets of larger historical systems. The holistic potential may not seem evident, however, since the immediate aim of most cross-cultural studies is the testing of specific hypotheses about how sociocultural variables are related.

The analytical coding of variables in comparative research is a labor intensive process. Opinion widely held several decades ago identified the independently chosen small sample as an efficient scientific method for extracting maximum benefits in terms of independent findings from an efficient (minimal: hence sampling) coding effort.

Cross-cultural studies based on haphazard (often mistaken for "random") sampling of a limited number of societies (e.g., fewer than 60) have begun to lose their appeal. Their deficiencies are now widely recognized: (1) low sample overlap with other studies, hence lack of cumulation in terms of our ability to study and interrelate a wide variety of topics on a common sample of cases, (2) insufficient number of cases to establish replication of findings in different regions, (3) insufficient number of cases for establishing connections among three or more variables at one time, (4) insufficient gain of ability to generalize to a larger universe since even "random" samples in this context are drawn from a constructed rather than natural universe; (5) invalidity of the claim that the cases, even in a small sample, are independent (hence: biased statistical estimates). Thus, when researchers spend months or years coding cases without the benefits of contributing to a cumulative database, the game may not be worth the candle.

Small random samples suffer the same deficiencies. Those that draw one case randomly (whatever the frame) from each stratum lose the sole benefit of true probability samples, the ability to compute standard errors. Random sampling by this means improves the representativeness of the sample over haphazard samples, but not necessarily vis-a-vis consciously constructed representative samples. They are a type of representative sample, and should be treated as such. It is often forgotten that probability samples must have within-stratum for valid statistical inferences. Thus true probability samples with two or more societies sampled within homogeneous strata have the added defect of a strong likelihood of selecting many pairs of non-independent cases.

George Peter Murdock began the trend toward cumulative databases based on representative sampling with his coded World Ethnographic Sample (1957), and Ethnographic Atlas (1962-80), which eventually extended to some 1,250 societies. Harold Driver (1957), focusing on North American Indians, began the trend of establishing cumulative regional databanks which are exhaustive rather than sampled.

Cumulative World Cultures databases are constructed of the following components: (1) a codebook of conceptually and operationally defined variables which records the manner in which societies have been coded; (2) coded data, generally consisting of a matrix of societies and variables with coded descriptive entries on each society for each variable; (3) a bibliography of primary and secondary ethnographic sources on each society, with indices to the sources and pages in (4) the ethnographic texts from which the coded information was drawn. Part I of this article discusses these aspects of constructing the cumulative databases published in the *World Cultures* electronic journal. Wherever possible, the discussion is phrased in terms of generic problems in social research.

Discussion of cross-cultural databases is in the plural, since cumulative databases will be published with different samples, sampling frames, regional and problem orientations.

The uses of cumulative databases include: (5) Continuous Area Data or comprehensive coverage; (6) Generalization from Sample Societies to Larger Frames; (7) Coding New Cases or Variables; (8) Mapping; and (9) Analysis. These are all potential uses for *World Cultures* datasets and are consequently discussed in Part II of this article.

2. DATABASE CONSTRUCTION

1. Codebook

Codebooks are guides to the computer coding of information that is recoded from published or unpublished variables coded on a set of societies by an original author. These articles are cited in the journal issues in which the codes are published in *World Cultures*.

A. Original Articles. Original articles by the authors of particular codes may be consulted to determine the precise definition of a variable, of coding categories, coding rules, underlying concepts, and attempts to establish reliability and validity of the codes. The codebook itself is a shorthand guide to labeling the variable and its coding categories or scales. In some cases it will be advisable to have both terse and expanded codebooks for the same database.

B. Recoding. The codebook for a computerized database usually provides numerical codes even if the author originally used other coding symbols (e.g., letters, short phrases, etc.). Quantitative indices, such as percent, temperature, size or number, are retained and not recoded into a smaller number of categories. Categorical, nominal, or ordinal codes are often reordered to form standard ordinal scales. An ordinal arrangement of codes is one where the assignment of numbers from low to high corresponds to some dimension underlying variation in the categories. Where different orderings of a variable are possible, only one is given in the database or codebook. The others can be obtained by recoding. This should be an option within the computer software for data analysis.

C. Editing the Codebook. A codebook should be clear and concise, ideally conveying information essential for defining the variable and coding categories, and allowing the original codes (if recoded) to be reconstructed. As these ideals are not always met, the codebook may go through several stages of editing. Publication of a codebook in *World Cultures* does not mean that improvements in editing the codebook will not be made subsequently. Readers are invited to submit improvements in the wording and format of presentation of the codebook.

D. Numbering of Variables. The easiest way to keep track of all the variables contributed by multiple authors of a cumulative database is to number the variables in a master list, with variables from particular studies in series. The order of studies and topics is arbitrary, and topical organization can be introduced by indexing.

E. Indexing the codebook. As a cumulative database grows to hundreds of variables, it requires indexing to keep track of all the ways that similar or overlapping topics have been coded. An index is arranged alphabetically by topic, giving variables or salient coding categories relevant to each topic. No standardized way of indexing codebooks has been developed, although the *Outline of Cultural Materials* (Murdock 1975, in 5th edition) might provide a uniform basis for indexing.

F. Inserting WARNINGS in the codebook about known deficiencies in the codes. The fact that a code is published does not mean that it is valid or reliable (see section 2). Where codes are discovered or suspected of excessive bias or unreliability, warnings are inserted into the codebook and explained more fully in the coding notebooks (see section 2G). For example, codes on sex ratios in adult or juvenile segments of pre-industrial populations are subject to rather severe biases if (1) females reach puberty before males,

(2) polygyny is practiced, or (3) girls marry younger on average than boys. These factors, widely occurring, tend to bias the ethnographer toward assigning young adults of the same age to an older cohort if they are female, and to a younger cohort if male. True ages are rarely known in preindustrial societies, and the resulting age biases skews estimates of sex ratios toward more adult females and fewer juvenile females than is in reality the case.

2. Data: Reliability, Bias, and Validity

Knowledge of reliability and validity of coded measures is crucial to the use and value of a cumulative database. Validity requires measurement of a properly identified or conceptualized variable (or coding category). A code may be reliable but not valid if it measures something other than what it is thought to measure. If a variable claims to measure A but in part reliably measures B rather than A, the B component is systematic bias. The unreliable component of measurement is random error rather than systematic. As is well known, unreliability lowers the expectation of correlation with another variable, while systematic bias does not.

A. Reliability: Multiple Codes and Measures. Because cross-cultural coding is labor intensive, authors of coded variables have rarely provided the double investment of labor costs to employ different coders to independently code each case, such as needed for an accurate measure of code reliability. Truly independent ratings of the same variable, however, are best left to independent research teams. With a cumulative database with hundreds of variables, it is often the case that variables of key interest to researchers will have been coded independently by more than one set of researchers. This is the case with the Standard Cross-Cultural Sample database for many of the central variables: subsistence mode and division of labor are coded independently in three separate studies; polygyny and form of marriage in four studies; and so forth for some major topics. No one has devoted systematic attention to questions of reliability. This is now possible with the cumulative database. Studies using variables on these central topics should either do preliminary studies of code reliabilities from multiple measures in the database, or report reliabilities if they are available from previous studies. In addition, it would help if authors of published codes would make available whatever information they have from their coding procedures and comparisons of independent coder ratings (often established for a subsample at an early stage of the coding operation) on the reliability of codes.

B. Validity: Face, Criterion, Construct, and Context. Validity is established in one of four ways. The most deceptively simple is face validity, or whether a variable seems to measure what it claims to measure. This is often ignored, but if the definitions and inferences which are used to code a variable are taken into account, much can be learned from an examination of face validity. J.W.M. Whiting, for example, conceives of cross-cultural codes in terms of high- and low-inference codes on the basis of face validity. In a low-inference code there is some directly observable behavior and few and simple steps

in moving from observation to interpretation in coding. In a high-inference code the variable is not a directly observable behavior but a construct inferred from a series of cues or symptoms, and the labeling of the code may be justified by theoretical inferences.

Criterion validity requires comparison with an established measure of the same concept. By contrast, Construct validity requires comparisons among measures of different concepts, through discrimination and/or convergence.

Context validity is the most subtle and most often ignored. It requires that the concept of a variable specify the range of contexts - including methods of measurement as well as settings of behavior - in which the variable is properly measured. Thus, variable X might be a valid measure in one cultural setting but not in another. Establishing context validity thus requires use of multiple methods (or contexts) of measurement as well as measures of multiple traits as a more general approach to construct validity (Campbell and Fiske 1959).

C. Bias: Levels of Systematic Error. There are at least four levels of potential systematic bias in comparative research, including (1) biased reports by ethnographer or informant, (2) biased construction of a sample or selection of bibliography, (3) biases in the way that codes are defined or that coders are making inferences, and (4) biases in the estimate of statistical parameters. In each case, the best thing to be done about known bias is to measure the source of systematic bias, and correct for it by some means (see Naroll 1965).

Ethnographer and Informant Bias. Raoul Naroll (1965) provides an excellent treatment of the problems of ethnographer and informant biases. For example, verbal reports of behavior by informants may be systematically biased as judged by their actual behaviors (selective memory, chunking, idealization, etc.). For ethnographers, shorter length of time in the field and lack of knowledge of the local language are known to produce biases in reporting of "insider" cultural knowledge such as magic, witchcraft beliefs, reproductive practices, intimate behavior, secrets, etc.

Sampling (and Bibliographic) Bias. Problems of sampling bias have led to the fiction of "random" cross-cultural sampling that is still in use for small-sample studies. Like a mail survey to 20,000 people with a 5% response rate, cross-culturalists are stuck with the fact that only a small fraction of the world's communities are well described ethnographically. Sampling biases are built into the cross-cultural enterprise by the types of communities that ethnographers have selected to study, that are unattractive to outsiders, that have shielded themselves, or that are so typical that they are passed over. There is no such thing as a "random" cross-cultural sample of human societies. One can follow Naroll's advice (1965), measuring possible sources of sampling bias and showing that these are not correlated with quality of information or more substantive variables so as to call into question the generality of particular findings. Or, one can use what variability exists with

available samples, preferably larger ones, to ascertain that a finding replicates in a large variety of maximally different contexts. Randomly sampling and resampling from a biased sampling frame may generate replicable results, but the effort is wasted since such replications do attack not the question of validity but only that of reliability.

Bibliographic bias is more amenable to control by the comparativist. This might result from undue reliance on English-language sources as opposed to those in French, Dutch, German, Spanish, Portuguese, etc.

Coding and Coder Bias. The way that a code is designed, worded, structured, or conceived, may be a source of systematic bias toward overestimation, underestimation, faulty inference, false analogy, etc. The instructions to coders may be a source of bias. The inferences used by coders, e.g., extrapolating from some assumptions of their own culture in making high inference codes, may be a source of bias.

Statistical Estimation Bias. It is well known that different sampling designs (e.g., cluster samples, stratified samples, random samples, maximal difference samples) have different effects on expected variances in the sample, and that the formulas for estimating expected variances (standard errors) ought to be adjusted accordingly. Estimates of standard errors, and precise adjustments to them, however, are possible only with true probability samples, where (1) every case can be assigned with precisely known non-zero probability selection in advance of sampling, and (2) every stratum has variance, i.e., more one case. In cross-cultural studies many groups have an expected probability of zero of being coded in the final sample because they have not been described.

This does not mean that we should not or cannot extrapolate findings to a larger universe of human societies, but simply that the standard errors of our estimates can only be guessed at, and many statistical tests have to be interpreted by additional rules of thumb.

One area where we do have more precise control over estimation bias is to correct the false assumption, underlying standard statistical tests, that the observations on the cases are independent, as in the tossing of coins. It is obvious that going from one society to the next does not redistribute values of variables on the model of a series of weighted-coin tosses. It is well known that statistical estimates of variance under a linked-observations constraint are systematically biased. Since interdependence or relatedness between cases -- linked by common history, common membership in larger political and economic systems, common origin, common language, or by trade, propinquity, intermarriage and the like -- is knowable and measurable, the common statistical tests must be adjusted, often severely, for the effects of linkage. Autocorrelation procedures provide unbiased estimation in the context of correlation and regression models. The autocorrelation problem is known to anthropologists and cross-culturalists as Galton's problem, after the 19th century statistician who elevated the problem to salience in cross-cultural surveys. Surprisingly, it is only geographers and cross-culturalists (and a few mathematical

sociologists) that have paid much attention to this problem, but it is one that besets all natural observation studies.

D. Quality or Bias Control Codes. Cumulative databases usually contain a number of bias control codes, ranging from the nature of evidence collected by and background of the ethnographers to the nature of the sample, the codes (e.g., hi/low inference), and the coders (experienced, trained, inexperienced). In some cases a good research design can utilize these control codes to test factors that threaten validity.

E. Autocorrelation and Estimation Validity. Cumulative databases will eventually contain a number of statistical bias control codes, particularly those relating to known connections or linkages between societies or observations.

F. Missing Data Codes. Standard missing data codes can be of service in a cumulative database. Since we want to minimize missing data, a distinction between partial information and inadequate information to judge is useful throughout. Use of the "." character for genuine missing information across all codes has been standardized for *World Cultures* databases. This is also the character recognized for missing data and case-exclusion in the SYSTAT and SAS statistical packages. By preserving the other 10 numerical digits (0-9) for information-carrying codes, including codes for partial information a few additional codes can be maintained as single column variables, which is helpful in condensing and mapping datasets.

G. Maintaining good database procedures. A number of database management procedures can help to maintain database quality. Access by the research community to the full range of coded data is an excellent way of helping to discover inadvertent errors: mislabeling of variables or codes, mistaken order of cases, or keystroke errors. Because these files, and the *World Cultures* journal itself, are maintained by an active cross-cultural research group, many errors of this sort have already been found and corrected. A high-quality database thus involves user responsibility for reporting discrepancies between data coded in the computer system and (1) the original ethnographic sources, or (2) earlier published sources, if such discrepancies cannot be attributed to a more recent updating of the information.

Errors of a substantive nature are those that contradict information in the primary sources. While codes ought to be preserved as originally published for purposes of measuring inter-coder reliability, it is also clear that improved substantive judgments ought to be used in amending codes where there is little ambiguity in the primary sources. If there are several independent codes of the same variables by different projects, it has been our practice to make substantive corrections to the best of the codes, and keep the other independent codes intact as originally published to facilitate reliability measurement.

In the present case, exercise over data quality is maintained by journal annotations in coding notebooks which accompany each set of codes and their codebook. For example, the first set of codes for the standard sample database published in issue 1#1 has a coding notebook (file STDCODE.1#1) to inventory all editorial changes and corrections to the coded data (but not to the codebooks). Careful justifications and annotations are given when changes are made to the codes developed by original authors.

Wherever possible, coded cross-cultural data should be keyed to sources and page numbers (sometimes to quotes) from the primary and secondary ethnography. *Development of the cross-cultural database will be greatly enhanced if authors of original codes will provide source and page numbers from their original coding notes wherever possible, but particularly for variables of theoretical concern to the research community.*

Given the phenomenal growth of computer data storage capacities, it is feasible and desirable to publish references to original source pages along with the coded data themselves. This has not been the practice to date in cross-cultural studies. This will prove to be invaluable a few short years hence when researchers will have access to the original ethnographic tests in computer-retrievable form at low cost, from personal computers.

It will also be helpful to assemble additional Quality Control Codes for the cumulative databases. The same may be said for intersocietal linkage codes.

Since the cumulative databases will be subject to continual cleaning and verification by users, the *World Cultures* journal policy is to make editorial changes in the database as required, record such changes in the annotated coding notebook, and provide free database updates to users available by returning the computer diskette for a particular issue to the editorial offices.

3. Bibliography

Each cumulative cross-cultural database is accompanied by a cumulative bibliography of ethnographic sources used in making coding judgments about the societies. Where possible, the bibliography is annotated to show which codes or studies used which sources in making coding judgments.

Many of the ethnographic sources for cross-cultural studies are foreign language materials provided in translation by the Human Relations Area Files. HRAF assembles and indexes a wealth of ethnographic material, on hundreds of societies, that it provides to member libraries as a worldwide service. Most (80%) of the Standard Cross-Cultural Sample societies, for example, have ethnographic material available in HRAF. For purposes of providing page number references to primary texts for the information on

which a coding judgment is based, it is expedient to index the sources for each society by the following bibliographic scheme:

A. Numeric codes (1-n) for ethnographic sources in the Human Relations Area Files as or 1986.
B. Letter codes (a-z, A-Z, aa-zz) for ethnographic sources not in the Human Relations Area Files.

4. Ethnographic Sources

Each ethnographic source relevant to coding information on a particular society needs to be evaluated in terms of its temporal and spatial or community focus. The correspondence between the source and the pinpointed time and focal group needs to be evaluated before the source is used or keyed to codes. This was done for the standard sample by Murdock and White (n.d.) and circulated to coders via bibliographic annotations.

In indexing codes to ethnographic sources it must be remembered that some discrepancies between sources are due to temporal, spatial, or other situational differences in the focus of observation. Thus, it is well to index the sources as to temporal and subgroup focus.

Bibliographic Quality Control variables eventually need to be coded for the important primary sources as part of a high-quality cumulative database. A relative quality-of-source ranking is developed by Murdock and White (n.d.) in unpublished bibliographies for the standard sample. It is improved upon by noting the rankings of the usefulness of sources by each of the coding projects that have contributed to the cumulative standard sample database.

3. DATABASE USAGE

5. Continuous Area Data

Continuous area datafiles provide relatively complete coverage on a region and all of its constituent societies that are ethnographically described, usually at one or more points in time. An example of such a database that is slated for publication in *World Cultures* (pending publisher agreements) is Jorgensen's (1980) *Western Indians: Comparative Environments, Languages, and Cultures of 172 Western American Indian Tribes*. A 443 variable computerized dataset codebook is published in the book along with maps of distributions of many of the variables. The dataset was computerized prior to the analyses reported in the book. Transfer to a microcomputer database thus represented a relatively simple task.

Murdock's (1962-80) *Ethnographic Atlas*, published in serial installments of the journal *Ethnology*, is also computerized. For many regions of the world it provides a quasi-continuous area mapping of cultural distributions on 93 variables. For other areas the coverage is spotty. Pinpointing data for ethnographic descriptions also vary considerably even within regions.

For a more literal ethnographic atlas showing the distribution of ethnic units circa 1961, see the section 9 on mapping.

The advantages of continuous area approaches, while labor and data intensive as opposed to representative sampling, may be enumerated briefly following the outline of my (1975: 300-310) review of the work of Harold E. Driver, a leading practitioner of this approach.

A. Ethnographic Context. Contextual factors can be better understood in the continuous area approach since the analyst can bring to bear a detailed knowledge of variables, historical context, interaction between societies, common origin, and regional systems.

B. Rival Hypotheses. Historical, evolutionary, and functional hypotheses as well as problems of inference (functional inferences from synchronic data, historical inferences from synchronic data, causal inferences from historical reconstructions) can be more easily assessed as rival hypotheses from the evidence of continuous area databases. Driver (e.g., 1956, 1966) excelled in identifying from synchronic or cultural survey data the different kinds of processes which operate side by side.

C. Assessment of Validity. The analysis of similarities between societies as well as correlations among variables can be more easily and validly interpreted in continuous area databases. Hypotheses about the interdependence of cases (through common origin, diffusion, trade, warfare, intermarriage, etc.) can be more easily assessed. Problems of identifying larger regional groupings or macro-cultural units are amenable to statistical classification procedures in this context.

D. Causal and Processual Inferences. By positing overlay causal and processual models (i.e., several simultaneously operating processes) in a continuous area context, mediating or third factors can be more intimately explored given knowledge of the region, and patterns of exceptions more easily identified for each given model. The realism that may be obtained in such models, while lacking the generality of working with a worldwide database, often exceeds that of studies based on cross-cultural samples. Inferences about temporal processes and time-lagged effects may be more easily assessed.

E. Replication. Continent-wide continuous area databases often contain sufficient heterogeneity and size to permit replication of findings within sub-regions. Continent-wide and regional replication of results is also an essential to testing findings for wider

cross-cultural studies. System and/or regional boundaries are also easier to identify both empirically and conceptually.

F. Identification of Units and Processes in a Fluid Field of Cultural Variation. Given the attention to larger and smaller units of cultural variation that is possible with a continuous area database, errors of misplaced concreteness are more easily avoided. Rather than assuming that "cultures" at the ethnic unit level are naturally discrete and internally homogeneous, the comparativist working with a continuous area database sees gradations, relative boundaries, and smaller units working within larger ones, or overlapping systems.

6. Generalization from Sample Societies to Larger Frames

Cross-cultural samples have both strengths and weaknesses compared to continuous area or comprehensive surveys. They save time in extensive compilation and allow greater investment in a deeper analysis of each case, or in terms of broader coverage of variables. They lose a good deal of information in terms of historical, regional, ecological, and world system contexts.

A. Reasons for Sampling. The purpose of a sample is, of course, to generalize patterns of relationship discovered in the sample to a larger framework by judicious selection of representatives. Thus, one should examine the larger sampling frame and methods of sampling construction in the light of this intended usage, rather than as simply a step in the construction of a database according to some formula. As discussed above, many of the questions of the validity of sample findings can be put to test using maximally different contexts within the sample as a means of establishing replication. To do so adequately, small samples are not sufficient.

B. Procedures of Sampling. Representative sampling by judgmental methods is in no way inferior for cross-cultural research than "random" sampling. I stressed in part I that true random sampling is nearly impossible in cross-cultural research.

C. Sampling Frames. Sampling frames are not always made explicit in cross-cultural research, although they should be. Murdock's (1957) World Ethnographic Sample, for example, is drawn from a universe of ethnographic descriptions that might be taken from a library shelf. Likewise Murdock and White's (1969) standard sample provided, as a product of Murdock's lifelong ethnographic reading, a selection of the best described societies in each of 186 cultural regions. But with every passing year the ethnographies, ethnohistories, and secondary analyses grow more numerous. A more systematic inventory of the available ethnographic sampling frame is needed. Building on the work of Murdock, Naroll, HRAF, and a host of other steps toward establishing a computerized ethnographic sampling frame will be published in *World Cultures*.

D. Pinpointing: Time and Focus. Each entry in a cross-cultural sampling frame, and each society sampled in a cumulative database, requires pinpointing in time and community (or other unit) focus for an assessment of the quality of sources. This task was begun by Murdock (1962-1980), Murdock and White (1969, n.d.), Naroll and Sipes (1973), and others, but this information requires assembly and compilation in computer readable form.

E. Sample Selection (& representation). In constructing samples for cumulative databases, it is advantageous to choose relatively independent well described societies rather than choosing randomly from the frame. Under random choice, many draws will have a paucity of information, and some pairs will be closely related societies. Both these types of draw are a waste of labor intensive coding time. The intermediate strategy of delimiting a small frame of well described societies and choosing randomly between them in a large number of strata demarcating culturally similar types is a valid and useful way of drawing a representative sample (basically no different than taking the best described in each stratum) but has little or nothing to recommend it in terms of statistical inference. In any case, the goal is representation of a wide variety of different sociocultural types rather than a random assortment.

F. Longitudinal Assessments. Observations are typically available on ethnographically well described societies at a variety of temporal foci. A cultural baseline sample such as the standard sample (Murdock and White 1969) explicitly draws the earliest of these temporal foci to minimize the effects of Western contact in the modern era, although such influences certainly cannot be avoided and must become an object of study as well. White and Burton's (n.d.) world systems codes for the standard sample provide longitudinal assessments of cultural contacts up to the early pinpointing date. This study has shown that a plethora of sociocultural change data are available not only from ethnographic sources, but from regional and economic histories. What is needed next is continuation of this timeline up to the present for the baseline sample. Continuous longitudinal changes (e.g., at one year or five year intervals) are usually not available for preindustrial societies. Records of known continuities or changes such as are available, however, can be extrapolated into time series, but precise methods for analyzing sporadic time sampled data need further work.

G. Source Identification. Problems of source identification and quality controls are discussed under points 3 and 4 above. With longitudinal assessment, intersocietal linkage data, and problems of embedding in larger systems, a host of new sources must be explored, assessed, mined for information, and inventoried into the bibliographic and source control system.

H. Sources of Sampling Bias. Problems of sampling bias are discussed under point 2B above.

I. Subsamples and Sample Expansion. Cumulative databases of moderate size provide some basis for analysis of focused subsamples dealing with specific sociocultural types, but often are inadequate. Special purpose samples (e.g., foragers, complex state societies, peasants) will need to be developed or incorporated. A standard set of baseline variables is needed to provide comparability between such subsamples and the larger cumulative databases, since it is not feasible to code subsamples on all of the extant variables in the cumulative database. Special subsamples can also provide a means for expansion of the larger cumulative databases, but careful attention is required to incorporate new cases while keeping an overall sample composition that is representative without undue weighting on a particular cultural type. Clearly, the (autocorrelation) methods for controlling for known relationships or linkages between sample units may be of use here, but remain to be explored.

J. Baseline Variables. Cumulative databases offer the potential of culling through multiple codings of conceptually similar variables, estimating reliabilities, establishing sources of invalidity, and compiling key indices on baseline variables that are of central theoretical importance. By reducing the number of variables in the baseline portion of the database, it is easier to expand the sample by new codings on only the baseline characteristics.

7. Coding New Cases or Variables

Advice on the construction of new codes will be provided in subsequent journal issues, along with further information on cross-cultural research techniques. A few salient issues touching on database construction and assessment of validity or reliability are considered here.

A. Indexing Codes to Ethnographic Sources. Perhaps the most important thing one can say regarding the coding of new cases or variables is that it is of utmost importance to the continued viability of a cumulative database to key the new codes to specific source and page numbers. A computerized index of page references for specific codes should be regarded as of equal importance as the coded information itself, as it is an indispensible index of the primary source of the information.

B. Use of HRAF. The Human Relations Area Files is an indispensable source of ethnographic information for cross-cultural research. A great number of foreign language sources are available in English only through HRAF translations. Although all of the HRAF materials on 360 societal units and over 5,000 sources (books and articles) are topically indexed using the OCM categories (*Outline of Cultural Materials*, Murdock 1975), a number of cautions must be observed. (1) The societal units are pinpointed only for the 60 society societies (Lagace 1977), with another 60 in process (Lagace 1979). Great care must be taken to insure that the HRAF sources indexed by OCM match the temporal and community or spatial foci of the particular sampling unit. (2) The HRAF

sources are not always complete for a particular pinpointed unit within the HRAF societal unit. Murdock and White (1969) provide an assessment of the quality of the HRAF file in coverage of the available sources for the standard sample. This is not a rating of the quality of the sources themselves. (3) The OCM indexing system does not always provide the best means of assessing the materials, since pages may appear out of context and out of sequence. They are excellent as a start for assembling page references, but must be supplemented by checking tables of contents and book indices, and consulting the original source (which is found in sequence in HRAF category 116).

8. Mapping

Cartographic display of sociocultural data is an extremely useful way of visualizing the complex patterns formed by regional systems, ecological systems, diffusion corridors, and other spatial aspects of interdependent societal data. *World Cultures*, through the database and the MAPTAB retrieval and mapping program (published in Volume 1, Number 3), provides a means for mapping the world distribution of variables in the cumulative databases. As new databases are developed (e.g., using different samples and variables), each will be compiled in a form that is amenable to analysis by the MAPTAB system.

The Laboratory of Ethnic Statistics and Cartography of the Miklukho-Maklai Ethnological Institute has compiled an Atlas Narodov Mira (Atlas of World Peoples) in Russian which shows the continuous area distribution of ethnic groups circa 1961. The only cultural data provided are language affiliation and population size of ethnic groups by country.

9. Analysis

This is brief review of issues in the construction of *World Cultures* databases, not a full treatment of methods for cultural analysis. Some salient themes emerge: we are looking for systems that operate at multiple levels in human behavior and meaning systems, which involves both the testing of concrete hypotheses about covariation, the positing of general processes, and the observation of patterns which may manifest themselves in different ways at different levels and in different contexts. Work in this fertile field may be exploratory and in the nature of pattern discovery as well as oriented toward hypothesis testing. Assessment of reliability and adequacy of measurement to fit the concepts employed is central. The nature of the categories and codes employed may become as much the object of scrutiny as the pattern that obtains in relationships between variables. We must not forget that we are observing societies that are linked historically into larger systems, and the analogy with the experiment, or use of statistics that assume independence of observational trials, may be inappropriate. New tools of analysis are available, and further developments will be stimulated by the availability of large scale databases to work with in which interdependencies among units as well as between

variables can be more rigorously and holistically assessed. Replication of findings, particularly under divergent conditions, and with a wide variety of conceptually equivalent measures, provides a key to establishing validity. More properly stated, we want to expose our findings to additional threats to validity to see if they withstand the test. A wide variety of data, tools for analysis, and scholarly resources contributed by an equal variety of authors, are suitable to the task.

4. REFERENCES

Barry, Herbert, III, and Alice Schlegel, eds.
1980 *Cross-Cultural Samples and Codes: Contributions from Ethnology.* Pittsburgh: University of Pittsburgh Press.

Campbell, Donald T., and D. W. Fiske
1959 Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56:81-105.

Human Relations Area Files
1976 *HRAF Source Bibliography: Cumulative*. New Haven: HRAF.

Jorgensen, Joseph G.
1980 *Western Indians: Comparative Environments, Languages, and Cultures of 172 Western American Indian Tribes*. San Francisco: W. H. Freeman and Company.

Laboratory of Ethnic Statistics and Cartography
1964 *Atlas Narodov Mira* (Atlas of World Peoples) [in Russian]. Moscow: Miklukho-Maklai Ethnological Institute.

Lagace, Robert
1977 *Sixty Cultures: A Guide to the HRAF Probability Sample Files (Part A)*. New Haven: HRAF.
1979 The HRAF Probability Sample: Retrospect and Prospect. *Behavior Science Research* 14:211-229.

Murdock, George P.
1957 World Ethnographic Sample. *American Anthropologist* 59:664-687.
1962-80 Ethnographic Atlas. *Ethnology*, issues 1-19.
1975 *Outline Of Cultural Materials.* New Haven: HRAF.

Murdock, George P., and Douglas R. White
1969 Standard Cross-Cultural Sample. *Ethnology* 8:329-369.
n.d. *Standard Cross-Cultural Sample Pinpointing Sheets and Bibliography*. Pittsburgh: Cumulative Cross-Cultural Coding Center.

Naroll, Raoul
1965 *Data Quality Control - A New Research Technique: Prolegomena to the Study of Cultural Stress*. New York: Free Press.

Naroll, Raoul, and Richard G. Sipes
1963 A Standard Ethnographic Sample, 2nd Edition. *Current Anthropology* 14:179-187.

White, Douglas R.
1975 Process, Statistics, and Anthropological Theory: An Appreciation of Harold E. Driver. *Reviews in Anthropology* 2:295-314.
White, Douglas R., and Michael L. Burton
n.d. World System Codes for the Standard Cross-Cultural Sample. NSF Proposal (preliminary draft in preparation).