October 1991

91-19

Working Papers in
Agricultural Economics

Rural Household Data Collection in Developing Countries:
Preparing the Data for Analysis

Tom Randolph

DEPARTMENT OF AGRICULTURAL ECONOMICS AND
CORNELL FOOD AND NUTRITION POLICY PROGRAM

It is the policy of Cornell University actively to support equality of educational and employment opportunity. No person shall be denied admission to any educational program or activity or be denied employment on the basis of any legally prohibited discrimination involving, but not limited to, such factors as race, color, creed, religion, national or ethnic origin, sex, age or handicap. The University is committed to the maintenance of affirmative action programs which will assure the continuation of such equality of opportunity.

RURAL HOUSEHOLD DATA COLLECTION IN DEVELOPING COUNTRIES:

PREPARING THE DATA FOR ANALYSIS

Tom Randolph*

---

* Presently an Agricultural Economist with the West African Rice Development Association and a Rockefeller Social Science Fellow (1992/93).

This paper benefited from comments by Ed Frongillo, Jerry Shively, Paul Higgins, and other collaborators of this data collection manual. Their help is gratefully acknowledged.

This paper was made possible by the cooperative efforts of Per Pinstrup-Andersen, Director of the Cornell Food and Nutrition Policy Program (CFNPP), and William Tomek, Chairman of the Department of Agricultural Economics, Cornell University. The author would like to thank both for financing the editing, production, and printing of this working paper. CFNPP funding was provided by the U.S. Agency for International Development.

The Department of Agricultural Economics at Cornell is concerned with economic issues in agriculture, natural resources and the environment, and rural communities. It is a unit of the College of Agriculture and Life Sciences. The Department strives for excellence in all functional areas (teaching, research, and extension), and departmental programs benefit from associations with other departments and programs on campus, such as CFNPP (see below). Working Papers are manuscripts which are subject to revision, and comments and suggestions are welcome. Nonetheless, these papers are intended as up-to-date reports of research and scholarly activities within the Department.

The Cornell Food and Nutrition Policy Program (CFNPP) was created in 1988 within the Division of Nutritional Sciences to undertake research, training, and technical assistance in food and nutrition policy with emphasis on developing countries. CFNPP has an advisory committee of faculty from the Division of Nutritional Sciences, College of Human Ecology, the Departments of City and Regional Planning and Rural Sociology; the Cornell Institute for International Food, Agriculture and Development; and the Department of Agricultural Economics and the International Agriculture Program. Graduate students and faculty from these units sometimes collaborate with CFNPP on specific projects. The CFNPP professional staff includes nutritionists, economists, and anthropologists.

Agricultural Economics Working Papers can be ordered from

Publications/Mail Room
Department of Agricultural Economics
52 Warren Hall
Cornell University
Ithaca, NY 14853
607-255-2101

The series of CFNPP Working Papers and this series of seven papers on collecting rural household data in developing countries also may be obtained by contacting

CFNPP Publications Department
1400 16th Street NW, Suite 420
Washington, DC 20036
202-822-6500

CONTENTS

ABSTRACT

This working paper series discusses a number of issues and techniques related to the management of data collected during a household survey. The data must be compiled, entered into computer files, and cleaned before they are in a format suitable for analysis. Ideally, data processing should serve to enhance the integrity and quality of the data. Attention is given to various types of errors that commonly characterize large survey data sets to help researchers plan strategies for identifying existing error and protecting the data from the introduction of further error. To guide the researcher, the steps involved in data preparation are described. Since computers increasingly are recognized to be indispensable tools in field research, their role in managing data effectively is highlighted.

FOREWORD

This paper is one in a series of seven working papers on collecting rural household data in developing countries. Between late 1986 and early 1988, six Ph.D. candidates from Cornell's Department of Agricultural Economics left to do the fieldwork in developing countries for their dissertations. Upon returning to Cornell in 1989, they discovered that they shared common experiences and frustrations while collecting household-level data for analyzing applied economic problems in developing countries. This series of working papers is the result of their collective effort to help other researchers avoid common pitfalls and build upon their experiences.

The working papers provide a practical field guide – for use together or separately – for individuals collecting a wide range of household information in developing countries. Each paper introduces the conceptual and practical difficulties involved in making different types of measurements or collecting different types of information. The guide is intended to provide readers with enough information about various methods so that those best suited to an individual's needs can be selected. Therefore, a variety of methods for collecting data are reviewed and the consequences of choosing one method or another are discussed.

Each working paper is organized into a section on conceptual issues, followed by a section on methods and organization. Conceptual issues address problems that researchers encounter when they move from a discipline's theory to empirical investigation. Often these include defining or measuring dynamic concepts or institutions such as the household, farm unit, time, or the valuation of goods. Related to this is evaluating whether or not to use certain variables in measuring rural lifestyles. In attempting to quantify particular aspects of rural economies, researchers realize that their definitions of selected variables do not always suit the reality of village economies. Thus, the sections on conceptual issues address the need to reconcile the researcher's theory and preconceived ideals with the realities of the survey site.

Although the related literature is reviewed in each working paper, the primary source of information has been the collective research experience of the authors. Examples of field experiences illustrate points made in each working paper. Many items that the authors felt they would have benefited from are included as well.

The target audiences are graduate students and other researchers, academicians, consultants, government employees, members of private voluntary organizations, etc., who are interested in collecting high quality socioeconomic, nutrition, and health data related to rural households in developing countries. In particular, the guide is for individuals who may not have had much prior experience in collecting this type of data, who may not have access to other current written material on data collection methods, or who may have some experience, but may not be aware of recent developments in data collection methodology.

One unique aspect of the series of working papers is its attempt to provide many examples of survey forms that have actually been used in field projects. Each working paper is built around the following question: How can survey forms and record keeping instruments be designed to assist the researcher in collecting high quality, nondistorted, less systematically error-filled data? Frequently, two or more forms that were used in different surveys (or in different rounds of the same survey) are discussed. The author has tried to be frank and honest, frequently providing criticisms of forms or tables that they used, but with which they failed to achieve the intended results.

Finally, a brief word on the use of 'he' and 'she' throughout the collection of working papers. Since the group of authors was equally divided into three men and three women, as a convention, generic third person pronouns and possessives (he, she, him, her) were consistent with the author's gender and should not be interpreted as a violation of political correctness.

The working paper series includes:

Paper Subject	Series Number	Author	Author's Country of Study*
Collecting General Household Information Data	91-13	Krishna B. Belbase	Nepal
Collecting Consumption and Expenditure Data	91-14	Carol Levin	Indonesia
Collecting Health and Nutrition Data	91-15	Jan Low	Northern Malawi
Collecting Time Allocation Data	91-16	Julie P. Leones	Philippines
Collecting Farm Production Data	91-17	Scott Rozelle	China
Collecting Off-Farm Income Data	91-18	Leones & Rozelle	Philippines, China
Preparing the Data for Analysis	91-19	Tom Randolph	Southern Malawi

* Each paper includes examples from other studies along with those from the author's country of study.

October 1991

Carol Levin and Scott Rozelle
Series Coordinators

1. INTRODUCTION

Once a survey instrument is designed and the desired information is collected, the data must be prepared for analysis. The final stage of the data collection process is often considered secondary and little forethought is given to planning the various tasks involved. While the researcher usually goes to great pains to ensure the collection of high-quality data, he may have only a vague idea of the steps necessary to subsequently transform those data from their raw form on questionnaires to variables suitable for statistical analysis. In retrospect, researchers often regret this lack of planning. By postponing the planning of data preparation activities until after data collection is well under way or finished, the researcher can grossly underestimate how much time – especially the researcher's own time – and other resources are required to complete data processing. "Data cleaning took a lot longer than I expected" is the researcher's oft-heard lament. More serious, though, is the possibility that poor handling may compromise the quality and integrity of the data. Without careful planning, the researcher may at best fail to capitalize on opportunities to correct erroneous data while still in the study area and may at worst introduce additional errors or lose valuable information during the process.

In the past, the neglect of proper planning for data preparation was perhaps attributable to technological constraints. Most data preparation activities had to be delayed until the researcher completed the survey and returned to the host institution where the necessary computer facilities were located. Data processing was, therefore, viewed as a postsurvey activity and was divorced from data collection activities both temporally and geographically. With the increasing availability of computer facilities and trained staff in most countries, and particularly with the development of portable microcomputers, data processing tasks are moving increasingly nearer the survey area, permitting integration with ongoing data collection. As a result, it is now more critical than ever that researchers recognize the importance of establishing a carefully designed data processing system during the survey's earliest stages. Only this way can researchers take advantage of opportunities to protect and possibly enhance the quality of data they have devoted so much effort in collecting.

This working paper series will help the researcher better plan for data processing activities by outlining the major steps entailed in preparing data for analysis and by describing the types of postcollection data problems that are typically encountered. Three assumptions underlie this discussion. First, we presume the data are generated from a survey or record-keeping system similar in structure and size to those described in the case studies. That is, the data set is large enough to be unmanageable by any means other than a computer-based system. Furthermore, hired personnel are needed to conduct the survey; data management is important in any size survey, but it becomes a particularly

critical issue when the researcher depends on others to help collect or handle data. Second, the survey instrument is presumed to be a questionnaire. While problems related to ethnographic data are not totally ignored, this discussion focuses on the more common short answer or coded response recorded on questionnaires. Third, recognizing that the computer is now an indispensable field research tool, we assume that the researcher has access to a computer while in the area where the survey is being conducted.

Postcollection data processing entails four steps. First, the completed questionnaires must flow from the enumerators in the survey area to some established collection point. As the questionnaires are transferred, they pass through a series of reviews that are designed to identify any errors in the data. If errors are caught at this point, it may still be possible to return them to the source for correction. Second, after the questionnaires are collected, data are entered into the computer. Third, the data are "cleaned." In this step, the data are carefully screened for errors and inconsistencies, and are then aggregated or transformed as needed to construct the desired variables for analysis. Fourth, the history and attributes of the data are recorded in a code book, thereby providing other interested researchers access to the data.

Before this series describes each step in more detail, it would be useful to review the types of data errors that typically beset computer-based data sets.

2. ERRORS IN DATA

Errors may be introduced into survey data from any number of sources at any point during data collection and processing. The researcher's objective is to minimize these errors either by identifying existing problems and correcting them or by preventing mistakes from entering the data in the first place. The researcher must know how to identify or anticipate errors and then how to deal with them. The error list below is certainly not exhaustive, but it covers many typical problems.

NONRESPONSE

Sometimes the enumerator fails to collect all of the desired data. The omission may be limited to a single question left blank because the respondent is unwilling to answer, or the enumerator may inadvertently skip a question during the interview. The omission may also entail whole interviews not conducted because of an enumerator's illness, absent respondents, unexpected holidays (e.g., a national period of mourning), or village events including funerals. In some cases, these errors can be corrected if caught in time. Particularly troublesome, though, is the case where weekly observations are being collected, but a complete week is missed. The researcher can improvise and try to gather recall data for the missed time period, but accuracy and comparability of such data are an issue. In other cases, it may be impossible to compensate for missing data. If a respondent is temporarily absent from the survey area and returns after the survey is completed, then no data can be collected, even retrospectively.

The problems created by nonresponse become apparent as the final variables are constructed. First, it becomes difficult to distinguish adequately between a null response and a nonresponse. Second, missing values can complicate aggregation tasks. For example, when the researcher is estimating annual expenditures from biweekly data for a household that was absent for two months, some type of ad hoc adjustment is required to derive the desired value. Keeping track of all such cases and making decisions about appropriate ad hoc adjustments are among the most tedious aspects of data cleaning. Finally, if a household has a nonresponse for even a single variable in the data set, the observation may be considered incomplete, and the possibility of using that household in the final analysis may be jeopardized.

MISPLACED DATA

Missing values can also be generated by information that was successfully collected but subsequently lost or "misplaced." The most obvious example is the case of a completed questionnaire that has been lost somewhere between the enumerator in the field and the data entry person in the survey office. Again, if caught in time, this error can be minimized by recollecting the data.

More subtle misplaced data problems also exist. When the enumerator records a response on a questionnaire, the answer may be condensed to fit the questionnaire format, and some information may be lost. This problem is especially apparent when answers are coded directly during the interview. Information can also be lost as the data are entered into the computer. For example, if a response on a questionnaire contains a brief description together with the corresponding code (i.e., farmer's time allocation is described as "uprooting and burning brush" and is coded as "12" for "field preparation"), and if only the code is put into the computer, then the detail included in the descriptive answer is lost. One disadvantage of losing the descriptive answer becomes apparent at a later stage when the researcher wishes to check the coded data for errors and finds there is nothing to check it against in the computer file. Of course, the researcher must weigh these disadvantages against the perhaps prohibitive cost of entering uncoded information into the computer.

INADMISSIBLE RESPONSE

A response is considered inadmissible if it lies beyond the range of acceptable values. For example, a coded answer of "8" would be inadmissible if the only codes defined for that question included "1," "2," "3," and "9." Inadmissible values usually result from recording errors, either by the enumerator during the interview or by the person entering the data. Other sources include the respondent's misunderstanding the question, which results in inappropriate responses, or improper conversions during variable construction (e.g., multiplying instead of dividing production by acreage to derive yield).

Inadmissible responses are generally easiest to identify, especially for numeric or coded data. Whether or not they can be corrected, though, depends on the availability of the correct information.

RECORDING ERROR

This category overlaps considerably with the preceding one. It includes primarily those mistakes made by the enumerator when recording the answer or performing computations or conversions on the questionnaire (what Casley and Lury [1987] refer to as a "slip of the pen," or in Malaŵi was called a "hand mistake"), and typographic errors made by the person putting the data into the computer. The result is, for example, a value of "100" reported by the respondent entered as "1,000" in the final data set. If the mistake yields an inadmissible response or a suspicious outlier when compared to the remaining

data, then it should be fairly easy to isolate. If not, then it may be almost impossible to detect.

INCONSISTENCIES

Inconsistent data refer to information that is collected independently during the survey and that proves contradictory. Contradictions can exist between parts of the same question on a questionnaire or between answers to the same question asked at different times or of different people. For example, when asked in separate interviews what field work was performed on the preceding day, the farmer responded that he and his wife planted maize for three hours, while the wife answered that, because of a funeral, no work was done. Inconsistencies also can be more subtle and may be evident only after the final variables are constructed, such as commonly found disparities between income and expenditures.

Of course, the dilemma is that the researcher does not know which of the two sources of information is correct. If the discrepancy is caught in time, additional information can be collected to help resolve the contradiction. Otherwise, the researcher must ignore the problem or attempt to reconcile the data using an ad hoc approach. Discovering data inconsistencies is particularly discouraging because it raises questions about the accuracy of the remaining data.

SYSTEMATIC ERROR

If an individual enumerator misunderstands the objective of a given question or if the enumerator conducts the interview or records responses in a manner different from other enumerators, systematic bias is introduced into the data. For example, if the questionnaire contains a query, "Who in the household has primary responsibility for a given field?" an enumerator's cultural bias may lead him to automatically assume the husband is responsible. Another enumerator may base his judgment on who does the most field work and may designate the wife as the principal operator. The solution is to prevent this error by carefully training and standardizing the enumerators and by pretesting the questionnaire. Even with the best of efforts, though, unexpected problems occur. Systematic error can also occur at the coding stage and even during data entry.

While some types of systematic error are readily apparent during initial data review, others are detected through comparison with other data. Certain patterns of error are discernible only over time or as a large chunk of data is reviewed at one sitting. This type of error may go undetected until the final analysis. If the researcher suspects systematic error exists in the data, it is usually possible to correct for bias during the analysis (i.e., in regression analysis using a dummy variable for observations associated with the suspect enumerator). The researcher is unable, though, to distinguish systematic error from true systematic differences between respondents, and this can weaken interpretation of the analytic results.

COMPUTER STATIC

This category contains a range of computer-related problems, which may be particularly acute if a computer is used in a survey area where the energy supply is unreliable. If the power fluctuates radically or is cut while data are being entered, the computer may leave "static" (for want of a better word) hidden in the middle of open data files. This problem is known to happen in dBase III Plus, for example. While some data are usually lost, many of the damaged files can be at least partially recovered.[1] "Static" can be troublesome particularly if it is discovered in a data file when the original questionnaires are no longer at hand. Other conditions, such as humidity and dust combined with little or no preventive maintenance, can also contribute to the potential for computer-damaged files.

Other computer-related problems are duplicate and blank records. Whether because of entry errors or glitches in the data entry program, unwelcome records always seem to work their way into data files. Blank records are easy to detect and are relatively benign. Partially blank and duplicate records are more problematic. These can be trickier to detect and, if left uncorrected, can cause bias (e.g., undercounting or double-counting) when data are aggregated to construct summary variables.

Four major points can be drawn from the preceding descriptions of data problems. First, data errors vary widely in the degree to which they can be identified or anticipated by the researcher. Errors run the gamut from the obvious to the indiscernible. Second, even if errors can be identified, data problems also vary in the degree to which they can be corrected or contained. In many cases, successfully correcting an error depends on a time factor, namely whether the problem is discovered in time to permit obtaining the additional information needed for correction. Third, errors also harm data quality and compromise the subsequent analysis to varying degrees. Fourth, the quest for error-free data incurs an ever-increasing cost at the margin. The researcher must take into account the first three themes mentioned above when weighing the advantages of further reducing error in the survey data versus the cost in survey resources, especially the researcher's own time.

The remainder of this series describes the steps involved in postcollection data processing and some possible strategies. Because much of the postcollection effort is devoted to maintaining and enhancing the quality of the data by minimizing errors as data are prepared for analysis, what applies to the individual researcher's own situation depends to a large extent on his personal assessment of the trade-off between error reduction and cost. Thus, the researcher must carefully evaluate the various sources and types of errors, including but not limited to those suggested above, that will likely arise during the survey.

---

[1] Comtech Publishing Ltd., P.O. Box 456, Pittsford, NY 14534-9990, Tel. (880) 456-7005, markets *dSalvage*, a dBase salvage utility that can help with most types of damage that occur to dBase files.

3. DATA COMPILATION

The first step in data processing is to collect the completed questionnaires and deliver them to where the data will be put into the computer. During this step, the questionnaires pass through a series of reviews to check for errors. Even before the survey begins, it is important to establish a schedule to ensure the smooth and timely flow of questionnaires. Without a system, questionnaires are returned haphazardly, opportunities for timely correction of errors are missed, and backlogs of questionnaires may arise.

Three elements to consider when planning such a system are (1) establishing a strict time schedule, (2) clarifying who will perform what checks, and (3) documenting the progress of questionnaires. If enumerators know when questionnaires must be submitted, they tend to stay on top of their work. Because checking questionnaires for mistakes can be very tedious, the discipline of a strict schedule helps to ensure that people with data-checking responsibilities do not lag behind. The schedule should be as rigorous as is feasible to facilitate rapid identification and timely correction of errors. It should also allow for returning questionnaires to the enumerator for correction, if necessary.

Each questionnaire should pass through a series of reviews before computer entry. At a minimum, at least two checks should be performed. The first is made by the enumerators themselves. They should know that they are responsible for examining each questionnaire for recording errors, for blank or incomplete responses, and for any inconsistencies before its submission. A supervisor or the researcher makes the second check. In addition to checking for the same types of errors as the enumerator, the reviewer evaluates overall quality of the data and checks for systematic error.

Checking the data helps minimize existing error. To limit the introduction of additional errors into the data during this step, survey personnel should maintain a log of questionnaires, which will guard against nonresponse and misplaced data. First, the researcher should construct a checklist of all households or respondents to be interviewed. That checklist should accompany each set of questionnaires (i.e., all the questionnaires for a single round for a single enumerator). The enumerator will record the date of interview and date of submission for each questionnaire, thereby ensuring that all interviews are conducted. *If an interview is not held, the reason should be recorded*; this information is essential later when the researcher decides how to deal with individual cases of nonresponse. A household that was not interviewed because of a death in the family, for example, is likely to be treated differently from one in which the household is harvesting its fields in a distant village. Questionnaires should move along the line of transmission in groups. Keeping the

questionnaires in plastic folders or bags protects them from physical damage – especially rain – while limiting the possibility that individual questionnaires will stray. Each person who handles the questionnaires will initial the list established by the enumerator and will double-check that no questionnaires are missing. In addition, each enumerator and data checker will keep a record of questionnaires that pass through their hands so, if needed, questionnaires can be traced. Generating the necessary forms and lists for such a system can be done easily on a computer.

When mistakes are discovered in questionnaires, and those questionnaires are returned for correction, it is advisable to keep each set together; individual questionnaires are easier to lose than an entire packet. Alternatively, enumerators could plan to visit the office periodically to correct questionnaires. Once corrected, questionnaires should be channeled through the standard set of checks a second time.

4. DATA ENTRY

Once the questionnaires are gathered and checked, the information they contain is put into the computer. As mentioned above, this function was formerly performed after completion of the survey and after the researcher's return to the home institution. Now, thanks to the microcomputer, data entry can be done while the researcher is still in the survey area. This means that in addition to managing data collection, the researcher is now responsible for simultaneously managing data entry. This section discusses some techniques of data entry and related issues that will help the researcher prepare for this task.

ADVANTAGES OF IN-COUNTRY DATA ENTRY

First, we emphasize the advantages of in-country data entry. The data entry person reviews all data as they are entered and identifies additional errors missed by the data checkers. If the delay between data collection and entry is short, it may still be possible to correct the error. Once the data are entered, the researcher begins cleaning the data, thereby identifying yet more errors.

Timely data entry can also permit the researcher to perform some preliminary data analysis. This prospect is particularly appealing because it opens up the possibility of an iterative process whereby preliminary results suggest refinements of established research hypotheses or even new hypotheses while the researcher is still in the study area and is still able to collect information. Unfortunately, experience indicates that these opportunities are actually quite limited; time-consuming data cleaning is required to prepare the data for analysis, and both the researcher's time and available computer resources may already be overtaxed. Despite good intentions, none of the authors were able to perform any significant analysis during their surveys.

Finally, it is usually cheaper to enter data while in the study area where labor costs are likely to be much lower than at the home institution. For large surveys, the savings earned may very well compensate for the additional expenses required to acquire and transport the computer(s) to the survey site.

CONSIDERATIONS

While planning and designing the survey, the researcher needs to consider a number of issues related to data entry.

Questionnaire Design

The first issue concerns the layout of the questionnaire itself. If the proper procedure is followed, the questionnaire data reflect what the researcher envisions as the various elements of the raw data set – nothing more, nothing less. To facilitate data entry, the researcher formats the questionnaire as much as possible to avoid confusing the person entering data. That person's eyes should not be forced to jump around the page searching for bits of data; responses should be clearly marked (i.e., coded answers in boxes separate from text) and laid out in an order that is easy for the eye to follow.

Descriptive Text

A second issue is whether all responses recorded on the questionnaire must be entered or just certain categories of data. Particularly relevant is the case of coded answers where the enumerator writes the original answer as descriptive text. For example, for a time allocation question, "took infant son to traditional healer" is recorded, and subsequently the answer is numerically coded as a medical care activity, "62." Should both the detailed description and the code be entered? Since it is likely that only the coded information will be used for the final analysis and that entering the descriptive text would slow down data entry considerably, it is tempting to ignore descriptive text. The researcher must recognize that by choosing this path, data are effectively lost. The additional information is valuable for two reasons. First, descriptive text in the computer data set permits cross-checking to ensure that the coded data are correct. Without descriptive text in the computer, the researcher cannot validate coded data, except possibly by visually checking the questionnaires. Time saved during data entry is then lost during data cleaning. Second, on-line descriptive information permits the researcher to return to the original descriptions to refine the codes, as needed. This information is particularly helpful for coding systems that include an "other" category. If this code is used frequently during the survey, the researcher may want to assign retrospectively new codes to common answers. The researcher should carefully consider the potential disadvantages of failing to enter all available information.

Data Aggregation

Another issue is the degree to which numeric data are aggregated before data entry instead of after. For example, two-week recall data are collected for daily hours of labor devoted to agricultural activities by a household, and the researcher is interested in using only the total in the analysis. The researcher can let the enumerators or data checkers total the hours over the two weeks and put only that final total into the computer. In the Philippines' survey, for example, the optimal approach was to limit data put in the computer to weekly totals that personnel computed by hand from daily time allocation records rather than entering the originally recorded data. Alternatively, researchers can put all raw data from the questionnaire into the computer and then program the computer to perform the desired operations. The trade-off is between saving time

in data entry and increasing physical storage (floppy diskettes) requirements on the one hand, versus more errors in computations and loss of information on the other. As with descriptive text, failure to enter detailed data can dramatically handicap the researcher's ability to subsequently clean the data and adjust the totals should the operational definition be modified. As a general rule, the researcher should place as much information as possible into the computer, even if incremental costs of entry time and diskettes are substantial.

Error Checking

When planning the data entry task, the researcher decides how much preliminary error checking to incorporate into the date entry process and how much to postpone until data cleaning. The decision is based primarily on the trade-off between investing resources in developing a customized data entry program that incorporates various error-checking functions, versus the increased time cost of post-entry data cleaning. Moreover, the researcher must consider whether postponing error checks will incur an additional cost in terms of timely identification of errors and their possible correction. In some cases, the original questionnaires are no longer available when data cleaning is performed, reducing yet further the possibility of verifying and correcting errors.

The types of error checks that can be performed during data entry, both for pre-existing error and error introduced by the person entering data, cover a wide range. The obvious and easiest checks to program are those for inadmissible or suspicious values. These errors are identified by comparing data to a predetermined range of validity, or by cross-checks for inconsistencies with related data. The only advantage to including many of these types of error checks in the data entry program is the possible benefit associated with early error identification.

At the other end of the range are error checks that require comparison of entered data with an outside data set. Tracking functions are examples of this type of error check. For instance, one of the first items entered from a questionnaire is the date the interview was conducted. When the date is entered, the data entry program could take that date and automatically compare it with information in a separate file containing a roster of all respondents. If it finds a date is already recorded for that respondent for the same round of questionnaires, it would return an error message; otherwise, it would record the new date in the roster. This procedure as two advantages: it avoids double-entering of questionnaires and, at the same time, maintains an up-to-date roster, which is used to quickly identify missing questionnaires.

Another example is an identification number verification. Each observation in a data set is typically associated with some type of identifier, such as a household number or field number, that allows the researcher to link the observation with other data for the same unit of analysis. If an error is committed when the identification code is entered, that observation creates havoc when the researcher begins analysis. A sophisticated data entry program asks for the identification code and then, after consulting a master list of codes,

displays the name associated with that code on the screen. That display allows the person entering data to compare it with the name recorded on the questionnaire. In addition, the program retains the identification code and automatically includes it with any of the records created for that questionnaire, thus avoiding further entry errors.

To minimize data entry errors, another strategy deserves mention: double punching. Double punching means entering the data twice. Data entry errors are identified by any discrepancies that arise between the separately entered values. Whether or not this is an option depends primarily on the relative cost of doubling data entry time versus alternatives for controlling entry error.

DESIGNING A DATA ENTRY SYSTEM

The basic components of a data entry system include the hardware, software, and a computer operator. For both the hardware and software, it is dangerous to make specific recommendations, given the fast pace at which technology is changing. Consequently, discussion of hardware is limited to a few generalities. For software, the focus is on the relative merits of options currently available.[2]

Hardware

If the researcher chooses the computer(s) for the survey, he should consider five characteristics for data entry. The first is dependability. Because service and parts for computers are usually unavailable locally, the researcher should get a machine with a reputation for low maintenance requirements. It may also be worthwhile to purchase a well-known, brand name computer, such as IBM or Olivetti, which is more likely to have an in-country service representative. No matter what the brand, the researcher should always run new equipment at least one week before transporting it to the study area; most mechanical problems with new computers happen in the first few hours of use.

Second, a hard disk is preferable, and the larger the storage capacity the better. Since data management and statistical programs typically involve continuous reading and writing from the hard disk, an important criterion to consider when selecting a system is the disk access speed, which is usually measured in milliseconds. Both a 3½-inch and a 5¼-inch floppy drive – preferably high-density – should be available to permit compatibility with other DOS systems

---

[2] The collective experience of the authors with respect to data entry and cleaning is limited to IBM-based systems, so all computer-related discussions in this series presume IBM-compatible hardware and software.

and software diskettes.[3] Minimizing the use of floppy disk drives is advisable in view of probable abuse from poor environmental conditions and inexperienced users. Problems associated with floppy diskettes are minimized to a certain degree by depending on the more durable 3½-inch diskettes rather than 5¼-inch diskettes. An attractive alternative is the Bernoulli technology, if the added cost can be managed.[4]

A third consideration is speed – the faster the better. Although faster running time probably has little directly observable effect on entry time, it can become a major factor when the data are subsequently cleaned. Even if increased speed translates into only a modest improvement in entry time, cumulative time savings over the complete survey can be considerable. Installing a math chip is advisable only if the computer is likely to be used to perform a substantial amount of statistical analysis.

A fourth consideration is including equipment to facilitate backing up procedures. Periodic – even daily – backing up of data files should be an integral part of data entry. Ideally, the researcher can minimize the time spent on backup as well as wear and tear on the computer, by installing a tape cassette drive directly in the computer.

Finally, a significant portion of information collected is likely to be numeric. Therefore, the researcher should use a keyboard with a numeric keypad to facilitate entry of numbers.

We should comment on Apple products. To the collective knowledge of the authors, there has been little experience with Apple products in field research to date.[5] Those products should not be ruled out, however, considering their general user friendliness, their portability, and their increasing compatibility

---

[3] If you have a high-density floppy drive, it is possible – and tempting – to format double-density diskettes as high-density (i.e., to format a double-density 3½-inch diskette, normally intended to carry 720K, to carry instead 1.4 meg). This procedure is **not** advisable since (1) there are often problems in reading the diskette on machines other than the one on which it was formatted; and (2) there is a much higher probability that the resolution of data definition on the media will deteriorate over time, resulting in corrupted data.

[4] A Bernoulli drive uses disks that combine both the moving parts of a conventional disk drive and the storage functions of a floppy disk. As a result, the drive itself cannot go bad (e.g., be knocked out of alignment) as is the case with a conventional disk drive; only the disks, easily replaced, can go bad. This is a valuable quality when computer service is not locally available. Also, individual Bernoulli disks have storage capacity comparable to hard disks, facilitating the transport of large data sets.

[5] An exception is Cornell University's Food and Nutrition Policy Program's current project in Guinea.

with IBM's DOS operating system. The major disadvantage is likely to be the lack of service representatives in many countries.

In the end, the researcher's budget and the anticipated size of the data set will primarily determine the final selection of a computer system. The researcher alone must decide what is appropriate in the context of his particular survey. Whatever system is chosen, the provision of a reliable power source to run the system is also important. In many cases, the computer is set up in a town that is in or near the survey site, and electricity is available. The computer must be capable, then, of switching to the available voltage; otherwise the researcher must add a suitable transformer to the system. The researcher should also include a line conditioner to protect the computer from power disruptions and, more importantly, from the large variability in current – both surges or overvoltages and brown-outs or sags – often experienced in developing countries. The best type of conditioner is an Uninterruptable Power System (UPS), which protects the computer from any external power problems while ensuring a supply of clean, steady voltage, and amperage.[6] If power is cut off without warning, the UPS can switch over to a temporary battery backup, thereby giving the researcher sufficient time to close any open files and turn off the computer. If electricity is not available at all, the researcher will need to provide a suitable battery or solar power system. The researcher should verify that the complete system works properly before taking it to the study area.

Not discussed here, but obviously indispensable is a printer.[7] Another periphery found to be quite useful is a small, hand-held vacuum for periodically cleaning the inside of the computer.

Software

Data entry programs span a range of levels of sophistication. The appropriate program for a given survey depends to a large extent on the expected size of the data set and how comfortable the researcher is with supporting the software when problems arise. In general, the larger the data set, the more sophistication is required, especially in terms of error checking. Unfortunately, in general, the more sophisticated a software package, the less it is user friendly. At the most basic level, a data entry "program" consists of simply typing the data directly into a word processing package without format limitations or error checks. For any but the smallest of data sets, this type of entry quickly becomes unmanageable.

---

[6] UPSs are periodically reviewed in computer magazines. Two companies known to manufacture UPSs that are suitable for foreign power systems are Sutton Designs Inc., 215 North Cayuga St., Ithaca, NY 14850, Tel. (607) 277-4301; and American Power Conversion Corporation, 350 Columbia St., P.O. Box 3723, Peace Dale, RI 02883, Tel. (800) 443-4519.

[7] In the Malaŵi surveys, dot matrix printers were found to be particularly useful for cutting stencils for questionnaire forms.

At the next level, data entry uses a standard spreadsheet, such as Lotus 1-2-3, which is menu driven and easy to learn. The spreadsheet can handle various types of data usually encountered in a survey (numeric, character, date, and logical) and permits data manipulation, especially numeric cross-checks and transformations. Other data-cleaning functions can be performed within 1-2-3, using 1-2-3 macros. Disadvantages include its limitations on the number of observations that can be stored in a single column in a work sheet, and memory problems with larger work sheets. A last disadvantage is its inability to customize the screen or data entry functions to meet the needs of the individual survey.

At an intermediate level are the quasi-related or nonrelational "flat file" database packages, such as Reflex, Q&A, and PCFile. These packages are generally menu driven and are very user friendly. However, they offer few if any error-checking facilities either as the data are entered or for subsequent data cleaning.

Compared to these, the SPSS Data Entry and dBase (III Plus and IV) programs are the Range Rovers of data entry.[8] Both permit data to be put in various formats: (1) in a spreadsheet (multiple observations listed vertically with individual fields displayed horizontally), (2) in a standard screen template by individual observation, or (3) in a customized screen template that can be made to resemble the questionnaire. In addition, both can be programmed to control the sequence of screens and error checks performed during and after data entry.

SPSS Data Entry is still quite user friendly, yet far more powerful than the intermediate-level packages. First, with a bit of very simple programming, the researcher can add conditional statements to instruct the program to skip sets of questions, to automatically fill in answers, or to branch to different sets of screens depending on the answer to a given question. Second, the researcher can define limits as to what are acceptable values for each individual field. For example, the values for a respondent's age can be restricted to between 0 and 100. A third feature is the functions for immediate preliminary data cleaning. The researcher defines a set of cross-checks to be performed and immediately identifies inconsistencies within a given data set. If a household member is coded as a student, for example, the program will check to make sure the person's age was in the 5- to 20-year range. Finally, with SPSS Data Entry, the data files are directly accessible for statistical analysis in SPSS-PC, which has data sets that have the nice feature of permitting variable descriptions and detailed value labels for individual codes. Other than its high price and a few minor bugs in its file management utilities, the major drawback of SPSS Data Entry

---

[8] SPSS Data Entry II is a product of SPSS Inc. dBase III Plus and dBase IV are Ashton-Tate products. Paradox and Foxbase Plus are alternative packages in this class. The authors have no experience with Paradox, so no comments are offered concerning its relative merits. Foxbase is a look-alike clone of dBase III Plus and even accepts dBase programs and data files without any conversion. Foxbase also appears to have some speed advantages compared to dBase, at least for some operations.

appears to be its ability to work in only one file at a time. This means that data from a questionnaire must all be entered into a single data file. If the questionnaire contains several sections that logically should be split into separate or independent files (i.e., household versus individual household member information, or crop production versus income information), the person entering data must repeatedly go through the same set of questionnaires, entering only the relevant section during each pass.

The dBase software can do all that SPSS Data Entry can and much more, but at the great expense of user friendliness. Although dBase can be menu driven, if it is to be used effectively, the researcher must learn the command-driven version. If the researcher knows some programming, he can design a very sophisticated data entry program. Besides customizing the screens to look like the questionnaire and setting restrictions on admissible values, the researcher can use a dBase program to work in several files, up to 10, at the same time. This permits the program not only to write data from a single screen to several different files, but also to look up information and to perform comparisons in reference files, as needed. The examples of tracking and identifier verification described in the earlier discussion of error checks can be performed in dBase, but not in SPSS Data Entry. Compatibility of dBase files with statistical programs is rarely a problem. If the program does not accept dBase files directly, then data in dBase files are first converted to ASCII format and subsequently transferred to the statistical program. Be warned, though, that dBase is somewhat fickle in how it treats blank numeric fields; it will sometimes automatically fill an empty numeric field with a zero, which cannot be easily removed or "blanked" out.[9]

Overall, dBase offers many advantages. For the typical researcher, though, dBase requires a substantial time investment to design a custom data entry program because effort is required to learn some rudiments of programming and dBase syntax. If the researcher's needs are more urgent or if the researcher does not feel comfortable with programming, then SPSS Data Entry offers a viable alternative. With SPSS Data Entry, the researcher can set up and run a data entry program within a day or two.

Computer Operator

Most surveys will require at least one full-time person to enter data as they are collected. Although desirable, hiring a person with extensive computer experience is certainly not necessary and, in typical developing country situations, is often not possible. Typing experience is, however, an essential skill. Lack of computer experience can be compensated for with a little bit of orientation and by setting data entry within a "closed" system. Here, the person entering data never goes into DOS; after turning the computer on, he is led from screen to screen with a series of menus. Such a system is easy to design using

---

[9] In dBase IV, a field can be "blanked," but still it must be done tediously cell by cell.

a program such as Automenu.[10] The operator simply chooses between entering the data entry program or performing some predesigned utility, such as backing up the data onto floppy diskettes or parking the hard drive for shutdown. The researcher can also protect the computer through measures such as disarming the Format command to avoid inadvertent reformatting of the hard disk.[11] Having a file recovery utility package, such as Norton Utilities or PC Tools, available when the inevitable disaster strikes (e.g., the person entering data accidently erases a directory full of data files) is also highly recommended.

If the person is not familiar with the computer keypad, particularly the numeric keypad, allow the operator to self-train with a computer typing program such as Typing Tutor IV.

For the sake of consistency, especially where more than one operator is entering data, it is important to anticipate and agree on a set of rules for various situations that may arise during data entry. For instance, does it make a difference whether letters are capitalized or not? Does a system of codes need to be established to differentiate between various types of nonresponses (not applicable, left blank/no answer, answer not in appropriate form, etc.)?

AN EXAMPLE

To illustrate what is involved with a data entry program, we will briefly describe the program used in the HIID Southern Malawi survey. We will give a step-by-step account of data entry for three-day recall of time allocation and labor use in the respondent's fields from a section of the questionnaire administered fortnightly to adult members of each sample household.

After the operator turns on the computer, the first screen displayed is a menu with data entry as one of the choices. After the operator chooses this option, dBase runs a set of hierarchical programs contained in two dozen or so program files. The master file creates a new menu that gives the choices of entering questionnaires from different components of the survey. After the operator picks the appropriate questionnaire, the computer initiates the input program. It immediately opens a number of files, indexes them, and links them automatically to each other. The screen graphics imitate the format at the top of the first page of the questionnaire, where the name and ID number of the

---

[10] Automenu is shareware produced by Magee Enterprises, 6577 Peachtree Industrial Boulevard, Norcross, GA 30092-3796.

[11] In older versions of DOS, if one made the mistake of typing "FORMAT" without designating the drive, DOS would reformat the hard drive! To prevent this from happening, the operator can rename the FORMAT.COM file that performs the formatting function by calling it something like KILLDISK.COM. The operator can then create a batch file named FORMAT.BAT, which will contain the command "KILLDISK A:". When the user now types "FORMAT," the batch file is executed, and only the A-drive will be affected.

respondent together with the date of the interview are recorded. The computer prompts the operator for the respondent's ID number. Once entered, the ID is checked against a master list of respondents, and the name corresponding to the ID number is displayed on the screen. The computer asks if the name is correct. If "n" is typed, the program loops back to the beginning; nothing is recorded in any of the data files (i.e., no harm done). The operator then sets the questionnaire aside for correction. If "y" is typed, the program continues. The program retains the ID number and automatically records it at the beginning of each record that it creates for this questionnaire. This function avoids recording errors in the ever-important identifier field.

Next, the operator enters the date of the interview. The program will not accept any dates from before or after the survey period. The computer finds the respondent's log of questionnaires in the master list and checks to see if a date has already been recorded for the time period covered by this round. If a date is recorded, the computer displays an error message, and the program loops back to the beginning for the next questionnaire, avoiding duplicate entries. If no date is recorded in the master list, the date of the current questionnaire is written to the file. (After entering all the questionnaires for this round, the operator prints the master list and can immediately spot any missing questionnaires.) At the same time, the program updates a separate file that keeps track of the number of questionnaires the operator entered each day.

Next, the program generates a new screen resembling the first section of the questionnaire where it records the respondent's activities for the preceding day in chronological order, with a line for each individual activity. Each observation includes spaces for the starting time, ending time, duration, brief descriptive text, and numeric code associated with the activity. The operator types in the starting and ending times (e.g., 6:00 and 9:30). As the operator enters the duration (e.g., 3.5 [hours]), the computer simultaneously calculates its own estimate of the duration from the times just entered and compares this to the figure entered from the questionnaire. If they are not equal, the computer asks the operator to double-check the numbers and it erases the observation, allowing reentry. This is an example of a cross-check that requires immediate correction of an internal inconsistency. After reentering the times and duration – which are double-checked once again – the operator enters the activity description and code. The computer writes these data as an observation in the time allocation data file, adding the respondent's ID number, the round number, the day of the recall period (e.g., Day "1"), and the corresponding day of the week derived from the date of interview. While this is done, the cursor moves down a line ready for the next activity to be entered. The previously recorded activities remain visible on the screen so the operator does not lose his place.

If any of the activities are coded as agricultural, the questionnaire should also note the ID number of the field where the activity was performed. If an agricultural code is entered, the computer prompts for the field ID number. The computer then simultaneously writes the record both to the time allocation data file and a field labor data file.

The program recognizes the last activity for that day of recall when it reads the latest time possible for an activity to end: 18:00. Once that observation is entered, the computer replaces the screen with one resembling the questionnaire's next section for field labor. First, the computer asks the operator, "Did any other people work in the respondent's fields that day?" If no, the program moves to the next section; otherwise, the program generates a blank observation for the operator to complete. After receiving the first observation, the program queries, "Are there any mistakes?" If yes, the operator can edit the observation before it is written to the data file. Once it is confirmed correct, the program records the observation with the same additional information as that generated for time allocation records. The program then asks, "Are there any more?" If yes, the program creates a new blank record on the next line of the screen below the still-visible records previously entered. If there are no more observations, the program moves on to the next day of the recall period.

The program continues to branch through the various sections of the questionnaire, skipping those that the operator notes as inapplicable. Once the questionnaire is completed, the program asks whether it should prepare for another questionnaire and re-initiates the loop. If not, the program returns to the main menu, where the operator chooses "Turn off the computer" to park the drive.

Unless the program crashes, the operator never deals directly with DOS or with the dBase command language. If there are any problems, the operator notes them in a log, recording the current questionnaire and a brief description of the problem. Even though the person entering data is not a trained computer operator, using a closed system for data entry requires minimum supervision. Management is limited to a daily or weekly review of the problem log. Even without direct supervision, it is possible to monitor the operator's performance in terms of speed and quality. Speed is measured by the number of questionnaires entered each day as recorded by the computer. Quality is assessed soon afterwards (one to four weeks) as preliminary cleaning and quality control programs screen the data. Although not done in the Malawi survey, the computer can be programmed to record each time the computer is turned on and off, so the operator's hours can be monitored as well.[12]

FILE MANAGEMENT

As the operator enters data, data files quickly grow. These files need to be routinely – and religiously – backed up. It is also advisable to devise a system for disaggregating the files into smaller, more manageable units. Otherwise, as data files get larger, the data entry program will become less efficient, and the program will appear to slow down. Insufficient storage space can also become a problem, especially when the operator tries to back up data

---

[12] Norton Utilities, for example, has a utility for timing execution of programs, and this could be captured and recorded to a file.

files onto floppy diskettes. Moreover, maintaining data in large files increases the amount of data that can be lost should a file be damaged.

A natural criterion for disaggregation is by unit of analysis. Nutritional data for children are kept in a file where the child's code is the identifier, while the mother's information is kept in a separate file using her code as the identifier. To link the information in the two files, all that is needed is a separate master file listing all of the children's codes and the associated mothers' codes. Files can also be disaggregated according to subject matter or section of the questionnaire. A file for income information, for example, is separated from expenditure data. Further disaggregation can be based on time of data collection, and the operator can create unique files for each survey round.

When creating the system of disaggregated files, the researcher should develop a simplified nomenclature for the file names. For example, the first three characters might represent the data category according to subject matter. The next character would identify the village number. Any following characters would be associated with the survey round in which the data were collected (e.g., L501, L502, and L503 for the first three rounds of labor data for Village #5; or I212 and E212 for the twelfth rounds of income and expenditure data for Cluster #2). The researcher should be sure to maintain the same structure for all file names. This system greatly facilitates repetitive tasks, such as combining the files or submitting them to data cleaning programs, which will include routines with DO loops written to perform the task automatically on all the targeted files. The researcher should keep file names as short as possible; there are always reasons to create subsets, and it is helpful if additional codes can simply be tacked onto the end of the current name. Disaggregating files dramatically increases the number of individual files; the researcher can create subdirectories for each file category to facilitate management.

5. DATA CLEANING

When the operator puts the data into computer files, they are ready to be transformed into variables for statistical analysis. This process is generally referred to as "cleaning" the data. The purpose of data cleaning is to identify and correct any remaining errors in the data, and to perform any conversions and aggregations necessary to transform the raw data into the desired final variables suitable for analysis.

Data cleaning comprises three rounds of cleaning activities. In the first, the researcher reviews the data for general quality and screens for error. This review can and should be performed on individual sections of the data immediately after entry. The second round of data cleaning is structured to deal with more subtle errors, which can be identified only after the complete data set is available. The end product of the second round is a "clean" set of raw data. In the third round, these raw data are transformed, using appropriate methods, into final variables.

Before discussing in more detail what each round of data cleaning entails, we must emphasize three points. First, the resources required for adequate data cleaning are often grossly underestimated, more than is the case for data compilation and data entry. While research budgets generally include line items for data compilation ("field supervision") and data entry, data cleaning is usually subsumed under the line item devoted to analysis, if there is one. Yet, even more resources, especially in terms of the researcher's own time, may be needed for cleaning the data than were required for putting them into the computer. The researcher should be aware of this and budget sufficient time after data entry for data cleaning tasks.

A second point is the need for consistency in data cleaning. It is generally preferable to subject all data to the same cleaning functions and to develop systematic rules when possible for any corrections or adjustments to the data, rather than to correct on an ad hoc, case-by-case basis. In practical terms, either (1) establish a checklist of commands to apply to the data or, equivalently, (2) use programs as much as possible through which all of the data pass as they are screened for errors, corrected, or transformed in any way. Otherwise, the researcher may end up using ever-varying criteria to edit the data.

Finally, the researcher should document any and all modifications made to the data as they are cleaned. In one sense, this task is simply an element of responsible research. More practically, it is in the researcher's interest to keep a log of how the data are adjusted and transformed so that the process can be reproduced if a variable needs to be reconstructed from scratch for some

reason – for instance, if the researcher's definition of net income changes slightly – or if the researcher suspects a mistake. In addition, documenting all of the steps facilitates writing code books for the data set.

ROUND 1: CHECKING ERRORS AND MAINTAINING QUALITY CONTROL

The sole purpose of the first round of data cleaning is to identify errors remaining in the data after they are put in the computer and, where possible, to correct those errors. Timeliness is again a consideration, and a timely first round of data cleaning is as important as timely data entry. Timely identification of errors increases the opportunities for higher quality corrections. For this reason, round one of data cleaning should parallel data entry, with data being cleaned as soon as possible. Careful thought must be given to round one of data cleaning tasks at the same time as the data entry program is designed.

Most of tasks in round one of data cleaning are simple error checks. The researcher screens the data files for errors missed or created during data entry. The researcher develops a checklist of error-screening tasks by reviewing the categories of errors that were listed at the beginning of this series. Nonresponse and misplaced data, for instance, result in missing or partially blank records. To catch missing records, the researcher generates a list of respondents from the data and compare these to a master list of all participants. To catch partially blank records, the researcher uses data management or statistical software to check the data file for unexpected blank fields. Then the researcher runs similar checks for recording mistakes, inadmissible responses, inconsistencies, and computer static. The researcher isolates and reviews outliers individually to see whether they are acceptable or should be returned to the field for verification.

The remaining round one task can be termed "quality control," one aspect of which is controlling the quality of data entry. To do this, the researcher takes a sample from the computer data file and compares it against the original questionnaires, noting the numbers and types of data entry errors discovered. Using the results, the researcher institutes corrective action by modifying the data entry programs or by retraining the data entry person. The researcher also effectively monitors the progress of the person entering data and can warn that person when the rate of error becomes intolerably high.

At the same time, the researcher also reviews the quality of the enumerators' performance and ultimately of the questionnaire. Both through comparison of a sample of questionnaires to the data files and through the frequencies and types of errors discovered during data cleaning, the researcher identifies systematic enumerator errors and unexpected problems with the questionnaire. Even if the researcher spends considerable time in the field directly supervising data collection, he may be unable to identify such problems until patterns can be discerned from a large amount of data. Lessons learned from data cleaning help refine the data collection efforts. As implied by the

discussion above, it is essential that the researcher personally perform the data cleaning if quality control is to be effective.

The researcher can ensure that all the data are cleaned in a consistent fashion by designing a computer program to combine all of the round one tasks. As a unit of data is put in the computer (e.g., all questionnaires for one round for one village), it is submitted to the cleaning program. The HIID Malawi survey used such a program, written in dBase. The researcher initiated the program by providing the survey round number and the village number. The program then located the appropriate files and ran a series of simple error checks, such as those described above. The researcher followed the progress of the program, noting the number of each type of error and, in some cases, correcting the data while still in the program. One part of the program forced the researcher to review data for three randomly selected questionnaires out of every 36 as quality control. Another part of the program generated rosters of respondents and dates of interviews so that missing questionnaires were readily apparent. Roughly three-quarters of an hour of round one data cleaning was required for every worker day of data entry.

ROUND 2: DEALING WITH MORE SUBTLE PROBLEMS

Some data cleaning tasks must be postponed until the survey is over and the complete data set is available. A complete data set is required because the researcher needs to review all available information to be able to identify errors that remain undetected during the initial screening, and to make certain corrections. Many activities in this round of data cleaning revolve around the researcher's ability to generate complete time series or inventories for each item for which multiple observations exist over the survey period.

As the researcher reviews each time series, errors not noticed during round one data cleaning may now be obvious. One example is inaccurate information about the relationship of individual household members to the head of household as recorded by the enumerator during the initial interview with the household. This error often occurs when the enumerator is treading on sensitive ground (i.e., a woman has children by different men or out of wedlock) or is confused by complicated family structures. If similar information is collected at the end of the survey when the enumerator knows the household better, then the two sources of information can be compared and the first set of answers corrected.

Another example is an inaccurate numeric value. Say, for example, a file contains monthly household livestock inventories, and for one month, it indicates a given household had four cows. During the initial screening, there would be no reason to suspect this value was incorrect. However, when the survey was over and the researcher used the computer data to generate a table of livestock holdings by month, it would be obvious that the value of four was probably a mistake because the household consistently reported owning only one cow for the remaining months of the survey. Upon review of the original questionnaire, the researcher would likely find that the enumerator had written the number "1" so that it was easily mistaken for a "4" by the person entering data. In some

cases, the researcher may have to visually review each time series individually to identify such errors. In other cases, it is possible to design simple programs that check for such inconsistencies. For some types of numeric data, it may be appropriate, for example, to write a program that tests whether each individual value in the time series is within a tolerable range – say, three standard deviations – from a computed moving average.

Another second round task is dealing with nonresponse and misplaced data. In some cases, the researcher may decide to estimate the missing data. Consider the above livestock inventory example. If data were missing for one round of the survey for this household, the researcher would likely feel justified in assuming that the inventory of cows remained constant at one and, thereby, would adjust the data accordingly. Such decisions can be made only through a review of the complete time series. In other cases, the researcher may wish to assign a code or special value (i.e., -99) for missing data to describe why the data are missing. The researcher performs such recoding at this stage of data cleaning. The researcher should record all such changes, including a brief description of each case.

Now the researcher may also run some final cross-checks for inconsistencies in the data. For example, if data are recorded both in the form of a descriptive text and in code, sorting the data by the coded value allows the researcher to visually review all descriptions recorded for each individual code value over the complete survey and to verify that they are consistent with the code's definition. Not only does this exercise permit detection of simple recording errors, but we find it surprising how code definitions, or enumerators' understanding of those definitions, sometimes "evolve" and change over the course of the survey unbeknown to the researcher. This is the only opportunity for the researcher to identify such problems.

Finally, the researcher may decide to revise some of the data definitions, particularly for coded data. If for a given data item the code representing "other answers" is used extensively in the collected data, then a refinement of the coding system may be warranted. By reviewing the answers included under the "other" code, the researcher can devise new code categories and recode the data, as necessary. In other cases, if certain code categories are little used, the researcher may decide to aggregate these under a single code and to eliminate superfluous codes.

We intend these examples of second round activities to be suggestive rather than exhaustive; the researcher may find additional tasks appropriate at this stage (depending on the characteristics of the specific data set) before concluding that these checks have satisfactorily minimized errors in the raw data.

ROUND 3: CONSTRUCTING VARIABLES FOR ANALYSIS

Finally, with the raw data "cleaned," the researcher begins transforming the data into variables for statistical analysis. In some cases, no transformation

may be required. The researcher can, for example, directly import the respondent's age into the data set of final variables and can link to it the appropriate respondent identifier. The bulk of data, however, will probably require some type of manipulation before it fits the appropriate variable definition.

We cannot overemphasize the necessity of documenting each step as the researcher performs these manipulations. All that is needed is a simple checklist of the steps taken and notes of any problems encountered. While transforming the data, the researcher must be ever vigilant against the possibility of corrupting the variables with new error. Performing an apparently straightforward operation on a set of data sometimes produces unexpected results for individual observations (e.g., taking the log of a number less than one). These mistakes may go undetected unless the researcher takes time to visually review the data after each manipulation. Another simple check is to verify that the expected numbers of cases were processed and observations were created during the operation. In addition, the researcher should perform the transformation by hand on one or two observations and compare the results with those generated by the computer to be sure the proper commands were used.

Activities involved in constructing variables generally fall into three categories. The first category is conversions required to transform data into the desired units. A simple example is changing land area from acres to hectares. A more complicated example is converting crop production from local volume units to its caloric equivalent. A farmer reports harvesting 15 baskets of maize still on the cob. First, the researcher must have an idea of the approximate volume of the baskets to estimate the total volume of harvested maize. Next, he needs a conversion factor to translate the volume of maize on the cob into an equivalent shelled volume. Then the researcher changes the volume of shelled maize to its kilogram equivalent. Finally, the researcher consults nutritional tables, if available, to convert kilograms of shelled maize into a caloric equivalent. As is obvious from the above example, the researcher often depends on supplementary information to perform the desired conversions. In some cases, supplementary information is available from published sources. More often than not, however, the researcher must collect the appropriate information during the survey. Clearly, the researcher must think through well ahead of time what conversions will be required and how they are to be performed if all of the necessary supplementary information is to be available when the variables are finally constructed. This need for planning is especially true for data related to agricultural production in areas where production is often measured in local units.

Aggregation is a second type of activity typically performed at this stage. The researcher aggregates data either within or across observations. For example, if an observation is constructed that contains all of the subtotals of expenditures by type of expenditure (e.g., food, travel, fertilizer, etc.), then summing the subtotals *within* the observation will yield an estimate of total expenditures. If, however, each observation represents a single expenditure transaction, then to obtain total expenditures for a household, the expenditure values must be summed *across* all of the records for that household. The same set

of error checks suggested above for conversions are applicable for operations involving aggregation. When aggregating data, there is an additional danger that some of the data being aggregated are not what the researcher thinks they are. For example, expenditures are summed across observations for each household, and one household is missing data for 2 out of the 12 months of the survey. Unless the omission is properly flagged, the researcher is likely to wrongly assume that the computed total for this household represents the annual total and is comparable to those for the other households. This type of error is very difficult to detect from computed totals. For this reason, it is important that the researcher understand the implications of using data containing any missing or zero values, and that these observations be identified, especially those missing values. The various lists and inventories generated during the preceding two rounds of data cleaning should be helpful in this respect.

A third activity is recoding. If, for example, instead of using an estimate of household income in the analysis, the researcher prefers to classify households by income quartile, then the income data are recoded from a continuous numeric value to a discrete numeric value of one, two, three, or four. Although recoding is generally a straightforward task, error checks are still appropriate.

The researcher completes variable construction using either a database or a statistical software package. Both will generally provide facilities for searching based on conditional statements, sorting, and report writing; thus both are roughly comparable. The choice depends primarily on the researcher's personal preference; he should work in whichever package he feels most comfortable or has had the most experience.

6. WRITING THE CODE BOOKS

When the final variables are constructed, one task remains: writing code books. Ideally, the researcher should write a first version of the code books as part of the pre-survey planning exercise. The researcher can develop the appropriate survey instrument by working backwards, detailing first the list of analytical variables needed to address the research question and then determining the list of field-level variables from which the final variables are to be constructed. Draft code books like these should guide the design of the questionnaire and the subsequent data processing.

Regardless of whether initial versions of code books are drafted before the survey, writing a final version of the code books is indispensable once the data are cleaned. First, code books serve as a researcher's reference when there are questions about the history or characteristics of a given variable file. Surveys usually generate multiple data files, and the researcher is likely to forget details about exactly how each individual file was set up and what steps were taken to construct the files. Second, well-written code books facilitate the exploitation of data by other researchers. There is no reason for a data set to be "retired" with the completion of the researcher's own analysis. The data set should be public and accessible to other researchers, especially in the country where the survey was conducted. If the researcher presents the data set to the government or to a research institution in the host country – an expectation of responsible research – then adequate documentation must accompany it.

A code book contains a detailed description of the structure and contents of the data file. These include (1) the attributes of the file, such as size, numbers of records, and file type (e.g., dBase or SAS); (2) a list of all the fields and their attributes, such as variable name, position in the file, size, and type (e.g., numeric, character, logical, or date); and (3) an explanation of what each variable represents. The explanation of the variable includes a brief description or operational definition and lists units (US dollars, kg), together with brief descriptions, where applicable, of individual values within that field, particularly codes.

A code book is not limited to lists of variable names and codes, however. The researcher should also provide some background on why and how the data file was created. This background briefly explains the purpose of collecting the data contained in the file. It also includes an outline of steps taken to generate the final variables: how the data were collected, what various data cleaning tasks were performed, and what manipulations were required to transform the raw data into values found in the data file. An assessment of the overall data quality and any warnings about potential problems are particularly useful. Given the nature of the information in the code book, it is crucial that the researcher

write the code book immediately after completing the variable construction process, while impressions about data quality and details related to the process are still fresh in his mind. Appendix A is an example of a code book from the Southern Malaŵi survey.

7. ETHNOGRAPHIC DATA

The nature of ethnographic data necessitates management techniques that are quite different from those for a statistical database. Ethnographic data are usually recorded either as entries in a journal, where a "record" may be a long description of the researcher's direct observation or perception of actions and conversations by the people under study, or may be verbatim transcripts of unstructured interviews with informants. Analysis proceeds directly from the initial recording of the information on paper; no data cleaning is required. The analysis depends on the particular technique or style adopted by the researcher, but typically involves a series of rounds to isolate and synthesize key information from the original observations, which are related to a specific theme or question of interest.

An ethnographic survey can generate reams of notes and transcripts. The objective of ethnographic data management is to make the relevant information in observations readily accessible to the researcher during analysis. Indexing is the principal technique used to facilitate data retrieval tasks. The researcher may develop his own indexing system or may adopt a preexisting system. The reader should see Werner and Schoepfle (1987) for an explanation and examples of indexing.

Storage, indexing, and cross-referencing of information in ethnographic records have traditionally involved some form of index card system. More recently, bibliographic software such as Notebook II has been developed to perform these tasks. Observations of variable length are put into one or more fields and are then indexed or searched on the basis of selected keywords. Before adopting a computerized technique, though, the researcher should carefully weigh the advantages of easier data retrieval against the probable high cost of entering extensive amounts of data.

8. CONCLUSION

This working paper series discussed a number of considerations and techniques related to the management of data collected during a household survey. The data must be compiled, placed in a computer, and cleaned before they are in a format suitable for analysis. The principal objective of data management is to minimize error in the data as they are processed. This series paid considerable attention to various types of errors that commonly characterize large survey data sets. It described strategies for identifying existing errors and protecting the data set from introduction of further error.

Throughout the discussion, we have stressed optimizing data quality. Well-designed data processing is necessary if the researcher is to maintain the quality and integrity of well-collected data. Clearly, as the various strategies outlined suggest, controlling error consumes research resources. Following all of the advice presented in this series requires substantial inputs of equipment, labor, and probably most important, the researcher's time. Obviously, trade-offs always exist, and the researcher must judge for himself to what degree additional resources devoted to error control justify the return in improved data quality. In a sense, we are preaching the ideal, and it is up to the researcher now to tailor the advice presented here to the characteristics of his particular survey and its available resources.

The researcher should recognize, though, that optimizing data quality is not solely a question of resources. Understanding the data preparation process and what types and where problems are likely to occur is half the battle. If the researcher can anticipate those problems, then he can plan carefully and can contain the problems at a relatively low cost. We hope this series provides the researcher with an understanding of the process, helps with anticipating common problems, and offers some practical hints on how to go about handling those problems. The rest is up to the researcher, who must use this information to plan and organize appropriate data management systems well ahead of time. In many cases, thoughtful planning before the survey can substitute for large amounts of resources that may be required to deal with unanticipated data management problems during and after the survey.

APPENDIX A

SAMPLE CODE BOOK

TOBACCO.DBF (dBASE III PLUS) DATA FILE

TOBACCO.DBF contains the information collected during the Tobacco Grower Survey, part of the HIID Malaŵi Research Project. The enumerators conducted one-time interview in July 1987 with those households in the sample that participated in the year-long agroeconomic and nutrition surveys and that were known to grow tobacco. The purpose was to provide an independent estimate of tobacco sales to compare to data collected on the monthly Income and Expenditures Survey, and to provide supplementary information about past tobacco-growing history for these "tobacco" households. The enumerator was to verify tobacco sales by examining the ADMARC card that each registered grower possesses and that records quota and sales information. Unfortunately, the survey was timed too late – the ADMARC cards had already been turned back to ADMARC to prepare for the next season.

All of the information collected during the survey is found in TOBACCO. Each record corresponds to a complete questionnaire. The 62 records represent 61 households. (Two tobacco growers live in Household 601, for which there are two records.) Information in the file includes type of tobacco cultivated (past five seasons), marketing outlet, whether or not the farmer is registered with ADMARC, tobacco quota (past five seasons), value of sales, and any additional comments recorded by the enumerator. The file underwent preliminary cleaning, including checking for blank or partially blank records and inadmissable code values, plus cross-checking across variables for inconsistent code values.

Please note:

(i) The enumerator recorded no quota information for Households 301, 302, 305, 311, 322, 326, 334, and 340. The quotas for these households are assigned the value of (-1).

(ii) There are two records for Household 601: one for the head, 60100, and one for his grandson, 60101. When summarizing information at the household level, the researcher must first combine the information from these two records into a single record representing the household.

(iii) Member 60100 had quotas for both Southern Division flue-cured and burley tobacco. It is possible to record only one of the two quotas as part of the observation; only the burley quota is recorded in TOBACCO. The quota for Southern Division flue-cured tobacco for 60100 is noted under the QUOTA field described below.

TOBACCO contains 62 records and occupies 9,412 bytes.

The database was created in dBase III Plus using the following structure:

Structure for database: A:tobacco.dbf
Number of data records: 62
Date of last update : 02/14/90

Field	Field Name	Type	Width	Dec
1	HHDID	Numeric	3	
2	MEMID	Numeric	5	
3	TYPE8687	Numeric	1	
4	TYPE8586	Numeric	1	
5	TYPE8485	Numeric	1	
6	TYPE8384	Numeric	1	
7	TYPE8283	Numeric	1	
8	MARKET	Numeric	1	
9	REGISTERED	Logical	1	
10	QUOTA8687	Numeric	4	
11	QUOTA8586	Numeric	4	
12	QUOTA8485	Numeric	4	
13	QUOTA8384	Numeric	4	
14	QUOTA8283	Numeric	4	
15	VALUE8687	Numeric	7	2
16	COMMENTS	Character	100	
** Total **			143	

Description of Individual Fields:

1. HHDID (n,3) Household identification code. Three-digit number "chh". The first digit "c" represents cluster number (1-6) and the remaining two digits "hh" represent the household number within the cluster (01-99).

2. MEMID (n,5) Member identification code. Five-digit number "chhmm". First three digits "chh" are the household identification number, HHDID. The last two digits "mm" represent the identification code for the household member. Males are numbered 00-09; females 10-19. The oldest household member has the lowest number in each range (i.e., 00 or 10); the number increases with decreasing age.

3.-7. TYPE8x8y (n,1) Type of tobacco cultivated during the 198x-8y season. Numeric code represents:
0 = no information recorded
1 = Southern Division fire-cured
2 = burley
3 = both Southern Division fire-cured and burley
4 = did not cultivate tobacco that season

8. MARKET (n,1) Marketing outlet. Farmers were asked where they had sold their tobacco production for the 1987 harvest. Numeric code represents:

0 = no information recorded
1 = ADMARC
2 = auction floor
3 = some at ADMARC, some on the auction floor
4 = to another smallholder farmer
5 = to a leasehold farmer or estate
6 = both at ADMARC and to a leaseholder
7 = to local traders
8 = in local markets
9 = did not sell tobacco because of poor yields (attributed to farmer being ill)

9. REGISTERED (L,1) Registered tobacco grower. Farmer was asked if he or she was currently registered with ADMARC.

.T. = yes, is registered
.F. = no, not registered

10.-14. QUOTA8x8y (n,4) Tobacco production quota for the 198x-8y season, measured in kilograms. Farmers are allotted a quota each season (for Southern Division fire-cured, the quota is allotted by ADMARC on the basis of the recommendation of local agricultural extension agents; for burley, no information on who allots the quota was collected). Unregistered farmers have no quota and typically market their production through another farmer. Note two problems:

(i) for Households 301, 302, 305, 311, 322, 326, 334, and 340, the enumerator failed to record any quota information – these records are assigned the value of minus one (-1);

(ii) Member ID 60100 reported having a quota for both burley and Southern Division fire-cured tobacco; in TOBACCO, only the burley quota is recorded (it was not possible to record both quotas on the same observation). The Southern Division flue-cured quota for 60100 is:

1986-87	1,200 kg
1985-86	1,200 kg
1984-85	1,000 kg
1983-84	1,000 kg
1982-83	900 kg

15. VALUE8687 (n,7,2) Value of tobacco production for the 1986-87 season, in Kwacha. Respondents reported cash receipts for their marketed tobacco production. No verification could be made since their ADMARC records had already been returned to ADMARC.

16. COMMENTS (c,100) Comments. Any relevant notes made by the enumerator on the questionnaire are reproduced.

REFERENCES

Berge, N., M. D. Ingle, and M. Hamilton. 1986. *Microcomputers in Development: A Manager's Guide*. West Hartford, CT: Kumarian Press.

Casley, D. J., and D. A. Lury. 1987. *Data Collection in Developing Countries*. Oxford: Clarendon Press.

Crawford, E. W., J. S. Holtzman, J. M. Staatz, C. Wolf, and M. T. Weber. 1988. "MSU Experience in Research Design and Data Processing/Analysis." Paper presented at the University of Zambia Food Security Research Network Workshop, Lusaka, Zambia.

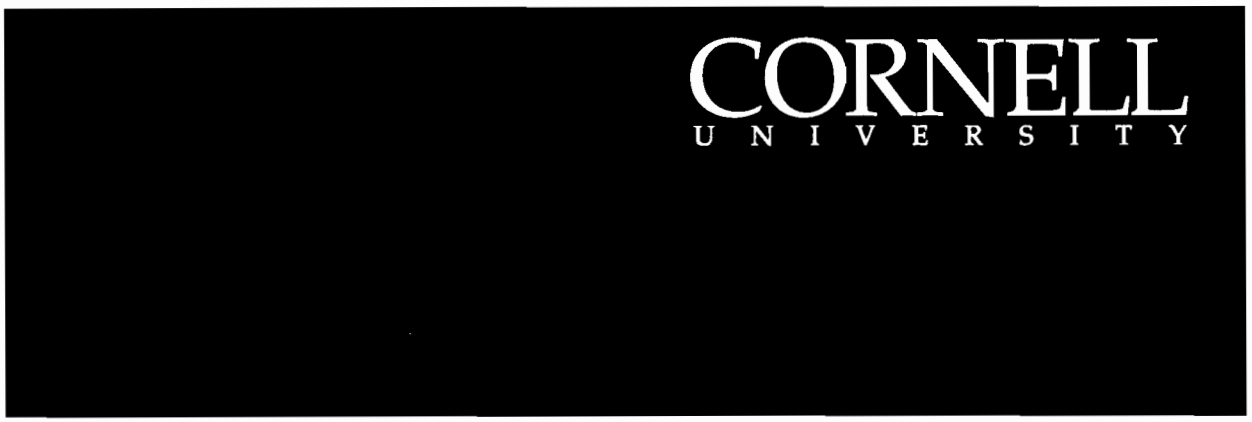
Werner, O., and G. M. Schoepfle. 1987. *Systematic Fieldwork, Vol. 2: Ethnographic Analysis and Data Management*. Newbury Park, NJ: Sage Publications Inc.

CFNPP WORKING PAPER SERIES

# 1	NUTRITIONAL STATUS IN GHANA AND ITS DETERMINANTS ISBN 1-56401-101-1	Harold Alderman
# 2	THE IMPACT OF EXPORT CROP PRODUCTION ON NUTRITIONAL STATUS IN COTE D'IVOIRE ISBN 1-56401-102-X	David Sahn
# 3	STRUCTURAL ADJUSTMENT AND RURAL SMALLHOLDER WELFARE: A COMPARATIVE ANALYSIS FROM SUB-SAHARAN AFRICA ISBN 1-56401-103-8	David Sahn & Alexander Sarris
# 4	A SOCIAL ACCOUNTING MATRIX FOR CAMEROON ISBN 1-56401-104-6	Madeleine Gauthier & Steven Kyle
# 5	THE USES AND LIMITATIONS OF INFORMATION IN THE IRINGA NUTRITION PROGRAM, TANZANIA ISBN 1-56401-105-4	David Pelletier
# 6	A SOCIAL ACCOUNTING MATRIX FOR MADAGASCAR: METHODOLOGY AND RESULTS ISBN 1-56401-106-2	Paul Dorosh et al.
# 6	UNE MATRICE DE COMPTABILITÉ SOCIALE POUR MADAGASCAR: MÉTHODOLOGIE ET RÉSULTATS ISBN 1-56401-200-X	Paul Dorosh et al.
# 7	DEVELOPING COUNTRIES IN SUGAR MARKETS ISBN 1-56401-107-0	Cathy Jabara & Alberto Valdes
# 8	MONETARY MANAGEMENT IN GHANA ISBN 1-56401-108-9	Stephen Younger
# 9	DEVELOPMENT THROUGH DUALISM? LAND TENURE, POLICY, AND POVERTY IN MALAWI ISBN 1-56401-109-7	David Sahn & Jehan Arulpragasam
# 10	PRICES AND MARKETS IN GHANA ISBN 1-56401-110-0	Harold Alderman & Gerald Shively
# 11	THE ECONOMICS OF CAIN AND ABEL: AGRO-PASTORAL PROPERTY RIGHTS IN THE SAHEL ISBN 1-56401-111-9	Rogier van den Brink et al.
# 12	COMPETITIVE ALLOCATION OF GLOBAL CREDIT CEILINGS ISBN 1-56401-112-7	Stephen D. Younger
# 13	AN ECONOMETRIC MODEL FOR MALAWI: MEASURING THE EFFECTS OF EXTERNAL SHOCKS AND POLICIES ISBN 1-56401-113-5	Yves van Frausum & David E. Sahn

Other Agricultural Economics Working Papers

No. 91-11	A Positive Theory of Agricultural Protection	Jo Swinnen
No. 91-12	What Role for *Leucaena Leucocephala* in Meeting Kenya's Fuelwood Demand?	Steven W. Stone Steven C. Kyle Jon M. Conrad
No. 91-13	Rural Household Data Collection in Developing Countries: Designing Instruments and Methods for Collecting Household Information Data	Krishna B. Belbase
No. 91-14	Rural Household Data Collection in Developing Countries: Designing Instruments and Methods for Collecting Consumption and Expenditure Data	Carol Levin
No. 91-15	Rural Household Data Collection in Developing Countries: Designing Instruments and Methods for Collecting Health and Nutrition Data	Jan Low
No. 91-16	Rural Household Data Collection in Developing Countries: Designing Instruments and Methods for Collecting Time Allocation Data	Julie P. Leones
No. 91-17	Rural Household Data Collection in Developing Countries: Designing Instruments and Methods for Collecting Farm Production Data	Scott Rozelle
No. 91-18	Rural Household Data Collection in Developing Countries: Designing Instruments and Methods for Collecting Off-farm Income Data	Julie P. Leones Scott Rozelle



Department of Agricultural Economics
New York State College of Agriculture and Life Sciences
Cornell University
Ithaca, New York 14853-7801
(607) 255-2191

Cornell Food and Nutrition Policy Program
308 Savage Hall
Cornell University
Ithaca, New York 14853
(607) 255-8093