PN-ABY-752

Family Health International

Women's Studies Project Guide to Study Documentation

May 1996

Family Health International is a non-profit research and technical assistance organization dedicated to contraceptive development, family planning, reproductive health and AIDS prevention around the world.

Begun in 1993, the Women's Studies Project aims to support social and behavioral science research on the immediate and long-term consequences for women of family planning programs and methods; and to help improve policies and programs through increased knowledge of the needs and perspectives of women.

Portions of this paper may be cited or quoted without permission for noncommercial purposes; appropriate credit should be given to Family Health International.

The Women's Studies Project is funded by the U.S. Agency for International Development (USAID), Office of Population, through a Cooperative Agreement (USAID/CCP-3060-A-00-3021-00). The views expressed in this paper do not necessarily reflect USAID policies.

Printed on recycled

Acknowledgments

Materials in this guide are compiled from 1) workshop materials on data management, documentation and verification conducted by the Family Health International Biostatistics Division for the Women's Studies Project and 2) existing standard operating procedures for project documentation and analysis verification at Family Health International. This guide was prepared under the direction of Nancy Williamson by Susan Pfannenschmidt, Emelita L. Wong and Cynthia Visness. Reference materials were prepared by Emelita L. Wong, Lucinda H. Glover and Rachel Disantostefano under the direction of Rosalie Dominik. Reviewers for this guide were Karen Hardee, Eilene Bisgrove and Debra Weiner. This work was supported by a Cooperative Agreement (#AID/CCP-3060-A-00-3021-00) with the U.S. Agency for International Development.

Contents and Uses of the Guide

This documentation guide consists of three parts: The Introduction presents an overview of the Women's Studies Project (WSP) at Family Health International (FHI) and the need for project documentation. Section I discusses study documentation, including the study operations manual (including the field operations manual and sampling design), data management procedures and preparation of the analysis plan. Section II defines the data analysis verification process. The focus of this guide is on documentation of selected processes in complex quantitative social science research. Researchers conducting studies with simpler designs and/or qualitative components should discuss with their FHI technical monitor how best to adapt the suggestions given to fit their studies.

We encourage researchers to document each phase of the project as it is being planned. Some researchers may feel it is easier to document study processes once a study is completed. However, delaying the documentation may result in loss of critical information. Personnel may not remember what they did, or why, even though it was perfectly obvious while the study was going on. Questions may arise, perhaps years later, when staff are involved in other projects or are no longer at their current jobs. Moreover, study documentation should be a helpful training tool for new project staff and a guide for current staff as each research step is implemented.

Most of the materials in this guide were taken from existing data management, verification and documentation procedures that FHI routinely uses for its papers and monographs or for reports prepared by FHI contractees. Since the WSP-funded research projects are conducted by local researchers (with technical assistance from FHI technical monitors), procedures will need to be adapted to local conditions and existing personnel resources. This guide serves as a reminder of the key research processes that need to be documented. It provides suggestions that in-country researchers and FHI technical monitors may refer to during the planning, monitoring and reporting of selected activities in the WSP research projects. Again, note that this guide applies mainly to *quantitative* studies, and we expect that it will be especially useful for large and complex studies.

Comments are welcome!

Comments on this guide are most welcome. We would like to hear the suggestions of researchers and technical monitors on whether this manual is useful and how it was adapted to local conditions. We all share the goal of ensuring high-quality data and analysis.

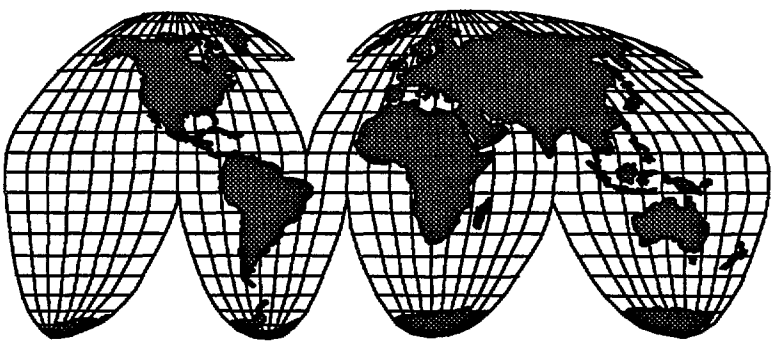


Table of Contents

Introduction **The Women's Studies Project** **4**

Section I. **Documentation**

A. **Study Operations Manual** **6**
What is a study operations manual and why do we need one?
What is contained in the study operations manual?
When is the study operations manual written and who writes and reviews it?

B. **Field Operations Manual** **8**
What is a field operations manual?

C. **Sampling** **11**
When is a sampling plan adequately documented?

D. **Data Management** **12**
How should data files be structured?
What is a codebook?
What are the usual steps in data management?
Figure 1. Two Methods of Computerized Data Collection

E. **Data Analysis** **18**
What is an analysis plan?
What is an analysis data set?
How is an analysis data set documented?

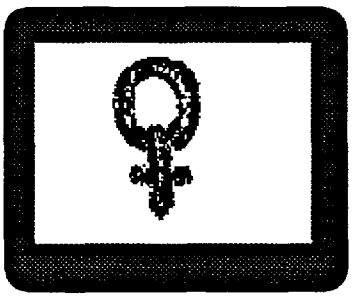
Section II. **Data Analysis Verification** **20**
What is data analysis verification?
When is verification conducted?
Whose responsibility is the verification process?
How does a researcher prepare for verification?
Are there any guidelines in the data verification process?
Study Closeout Procedures

Glossary of Terms **27**

Appendix A **Naming Variables** **29**
Appendix B **Creating Consistency Checks** **31**
Appendix C **Example of a Log Sheet** **33**
Appendix D **Example of a Sampling Plan** **34**

Introduction

The Women's Studies Project



The Women's Studies Project is being implemented by the Women's Studies Division of Family Health International and is funded through a Cooperative Agreement with the United States Agency for International Development. The purposes of the Project are 1) to support social and behavioral science research on the immediate and long-term consequences for women of family planning programs and methods and 2) to help improve family planning and related reproductive health policies and programs through increased knowledge of the needs and perspectives of women. The WSP is supporting more than 20 research projects in a variety of developing countries. The emphasis countries include Bolivia, Brazil, Egypt, Indonesia, the Philippines and Zimbabwe, where the WSP supports multiple full research projects, including data collection and analysis. In these emphasis countries, in-country advisory committees are set up to advise researchers on the various aspects of the research study. The WSP is supporting single studies in Mali and Jamaica, and has supported secondary analyses of data from the Philippines, Bangladesh, Malaysia and Nigeria.

Subject Matter

The WSP looks beyond the determinants of family planning use and its effects on fertility. Instead, the WSP is investigating how family planning affects women at different stages of their lives, and in their multiple roles as individuals, wives, mothers, workers and participants in public life. The WSP is documenting what women themselves say are the positive and negative effects of contraceptive use on the various aspects of their lives.

Methodology

The WSP is assessing family planning use and its association with various aspects of women's lives from both users' perspectives and through objective indicators. Both qualitative and quantitative methods are employed in WSP studies to capture women's perspectives.

Local Ownership and Informing Policy

In each emphasis country, the WSP is bringing together three groups: family planning professionals, women's groups and researchers. Representatives of these groups are serving on the in-country technical advisory groups that make recommendations on research priorities, women's groups' participation, informed consent and confidentiality procedures, information dissemination and policy impact. In bringing these groups together early, the WSP is encouraging local ownership of the research studies, providing direction for dissemination of research results and establishing linkages between groups that are interested in improving family programs and the lives of women.

The WSP has encouraged all phases of the research projects to be locally initiated, with research project monitoring shared by FHI technical monitors as well as local WSP secretariats. FHI staff and consultants provide technical assistance to in-country colleagues on each stage of their research to ensure that high-quality studies are conducted.

High-quality data and documentation

Both in-country and FHI researchers aim to provide high-quality research results of use to the various stakeholders: to family program professionals, who are involved in delivery of services as providers; to researchers, who are currently or may be encouraged to do research on family planning and women's issues; to women's groups, who can use WSP research results in their advocacy and intervention projects; and to policy-makers, who can translate research results into policy change.

The WSP is encouraging various procedures to ensure high-quality data in all the research projects it is supporting. Data quality procedures are outlined in the project agreements prepared jointly by in-country researchers and FHI technical monitors. As part of a continuing effort to ensure high-quality data, this guide on WSP documentation has been prepared to ensure an adequate and uniform documentation of key research processes. Suggestions on what should be documented, as well as other aspects of the research process, are presented.

The Women's Studies Project

- Consequences rather than causes of FP use
- Qualitative as well as quantitative methodologies
- Partnership between researchers, women's groups and family planning professionals

Section I

A. Study Operations Manual

Q. What is a study operations manual and why do we need one?

A. A study operations manual describes each step in the research process. It contains most of the elements included in the WSP subagreement but goes beyond the subagreement. The subagreement is a contractual document containing an overview of the project and defining the administrative and legal links between country researchers and FHI, whereas the study operations manual describes in greater detail how each research process is implemented. The purpose of the study operations manual is to document how the study was conducted and how typical problems (for example, nonresponse) were handled. Ideally, plans for each research step are written down in the operations manual prior to the implementation of the project. Some researchers may be more used to a "do now and document later" approach. While we realize that some aspects of a research study lend themselves to such an approach, we ***do not*** recommend this strategy, since as time passes, the researcher risks forgetting details that are important to include in the manual.

Q. What is contained in the study operations manual?

A. The study operations manual should include, in well defined sections:

- ☐ clearly stated objectives and clearly defined hypotheses
- ☐ a description of the study population and number of groups, including inclusion and exclusion criteria
- ☐ sample size and justification
- ☐ sampling plan and recruitment procedures
- ☐ instrument development procedures and final instruments
- ☐ procedures for obtaining informed consent and for maintaining confidentiality
- ☐ procedures for handling subject loss and nonresponse
- ☐ methods and timing of data collection
- ☐ procedures for handling errors in study procedures
- ☐ data management and analysis plans
- ☐ training manual for supervisors and field team
- ☐ a copy of each questionnaire and item-by-item instructions
- ☐ data analysis plan
- ☐ study closeout procedures

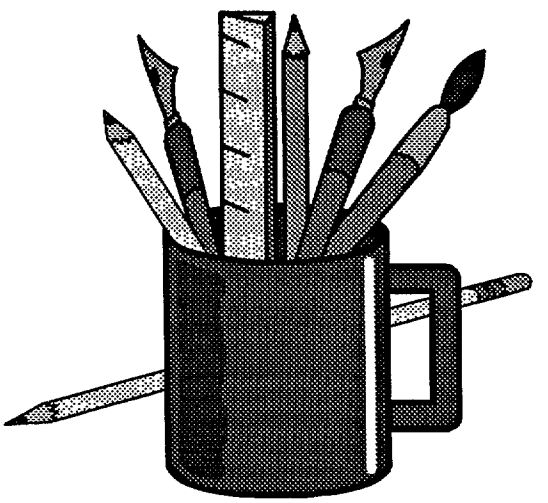
Certain elements of the study operations manual will also comprise the field operations manual. (See below.)

Q. When is the study operations manual written and who writes and reviews it?

A. We recommend that the operations manual be prepared before project initiation. Having to write the manual is one of the best ways to thoroughly think through all study procedures and to start off "on the right foot." The principal investigator (PI), project coordinator or other staff may write the entire manual, or each may write parts and combine them for the final product. While it is preferable that the entire plan be written before the start of the study, when this is not possible, a plan for each specific research step should be written prior to its implementation. As each step is implemented, the researcher should write an evaluation of that step in the manual for future reference. Documentation needs to include any deviations from the plan, and the reasons for such deviations.

The PI retains the main responsibility for the research plan and its implementation, evaluation and documentation. The FHI technical monitor or project consultant would be happy to review the plan and make suggestions for improvement. The in-country researchers and FHI technical monitor typically prepare a checklist and schedule of technical assistance and review for each aspect of the study. A schedule of the FHI review should be agreed upon by both the in-country researchers and technical monitors.

At project completion, in-country researchers submit a draft report to the FHI technical monitor, which should include sections on research procedures. The technical monitor then will work with the PI to prepare a final project report to be sent to USAID.



B. Field Operations Manual

Q. What is a field operations manual?

A. The field operations manual is the part of the study operations manual that is directed specifically to the staff who collect the data. The field operations manual should state the study objectives and identify the roles of all field staff.

Q. What is a field operations manual?

A. The field operations manual should provide specific instructions regarding all study procedures, including:

- ☐ training, including gender sensitivity training
- ☐ operational definitions, such as who is an eligible respondent
- ☐ question-by-question specifications
- ☐ how to approach people to participate in the study
- ☐ informed consent and confidentiality procedures
- ☐ how to answer participants' questions
- ☐ when and how to refer participants to local resources
- ☐ how recruitment or sampling is to be carried out
- ☐ how interviews should be conducted
- ☐ how to record changes or corrections to data
- ☐ what to do in case there are problems in the field
- ☐ whom to contact with questions
- ☐ instructions for special data handling, such as linking interview questions with qualitative data

Examples of completed questionnaires are also helpful, especially if there are complicated skip patterns. The field operations manual also includes the respondent tracking system.

Tracking System

A tracking system involves recording the status of each respondent from selection up to the point where her or his data are processed and considered part of the analysis data set. The tracking system includes the "status of events" associated with the completion of the interview and the processing of the completed instruments. In studies where the participants are to be interviewed at home, the result(s) of the home visits must be specified using result codes. The interview

result coding system will vary from project to project but should be comprehensive enough to differentiate eligible from ineligible participants, and, in cases where the interview is not completed, the reasons for this. Temporary codes are assigned until the interviewer and field supervisor agree on a final code.

Tracking System Coding Suggestions

- Interview is completed with eligible respondent.
- Interview is completed by eligible participant with assistance from other household member.
- Eligible respondent declined to be interviewed.
- Eligible respondent is not at home.
- Eligible respondent is incapacitated.
- Interview is completed by ineligible respondent
- No one in household is found eligible based on the criteria for selecting respondents.
- House is vacant or no one is home.

Longitudinal studies usually have a similar kind of tracking system, but should indicate the range of dates when a given visit may take place, as well as when the visit was actually made. A handwritten log sheet or logbook will suffice to document this information.

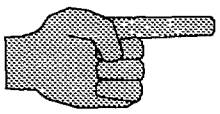
For respondents identified through an area household survey, include the following in the tracking system:

- sampling variables (strata, primary selection unit, housing unit)
- Number and dates of visits for screening (maximum needs to be determined by principal investigator)
- result codes for each visit

For those households successfully screened include:

- presence of eligible sample persons
- indicator for selection of eligible sample persons
- number and dates of visits for interview of sample person
- result codes for interview at each visit of sample person
- final result code for the interview

Events associated with the processing of the instruments from field editing to verified data entry are also part of the tracking system. Below are the events associated with processing the instruments/records.



See Appendix C for an example of a tracking log sheet.

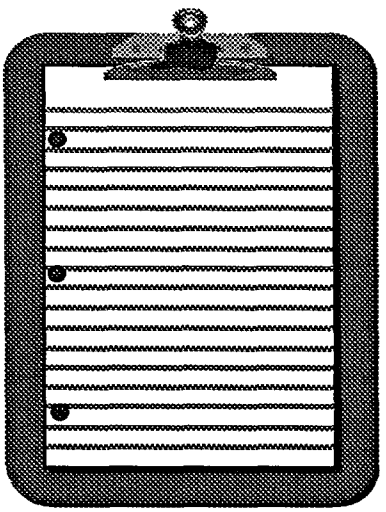
Records and Forms

Once data are collected, the forms should also be tracked. A log sheet, or other tracking system, should be developed to maintain information on when each data form (or batch of forms) was:

- ☐ edited by the field supervisor
- ☐ sent for keying
- ☐ keyed/double keyed
- ☐ loaded into the database (if this is a separate step from data entry),
- ☐ checked
- ☐ queried
- ☐ changed
- ☐ filed

Part of the tracking system is the code for the interviewer assigned to the respondent and the interviewer's field supervisor. Decision rules on assigning the correct status for each respondent also comprise the tracking system. In general, the field supervisor jointly with the interviewer assigns the final interview result code for a respondent and the data manager or data clerk assigns the records/form event status.

In smaller studies, tracking codes can be recorded directly onto the data collection instruments. In larger studies, however, a database -- either manual or electronic -- needs to be created for recording the status of each interview attempt.



C. Sampling

Q. When is a sampling plan adequately documented?

A. A sampling plan is adequately documented when someone with a minimal knowledge of sampling procedures would be able to select a sample correctly if s/he were given instructions and access to the mechanism or list for drawing the sample, and when a data analyst then would be able to correctly calculate sampling weights.

Sampling plan documentation should include:

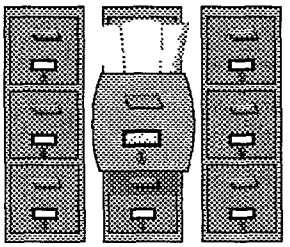
- study objectives for the sample
- dimensions of the target population and the sampling frame
- sample size and rationale
- selection procedures, including:
  - stages of selection
  - stratification
  - units of selection and sampling frame at each stage
  - clustering
  - implementation of randomized selection
- estimation procedures, including:
  - construction of basic analysis weight (see below)
  - adjustment of analysis weight for nonresponse undercoverage and overcoverage, multiplicity of inclusion
  - estimation of variance
  - stages of selection

Sampling design documentation is needed for studies with household area survey components. For such studies, the original and adjusted probability of selection of the sampling units must be contained in the data analysis file, and specific instructions on how these probabilities are combined must be included in the sampling plan. The resulting analysis weight must be contained in the data analysis file. A sample sampling plan can be found in Appendix D.

D. Data Management

By following these conventions and procedures, data management and analysis will be made easier:

File Structure

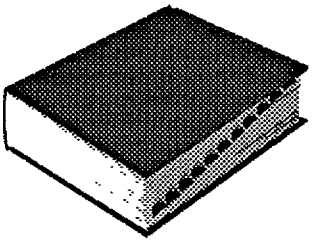


Q. How should data files be structured?

A. The following steps should be taken when structuring data files:

- ☐ Each unique questionnaire or form should be a separate data file.
- ☐ Each questionnaire and data file should include identification (ID) variables so that each individual form can be uniquely identified. Each participant should have an identification code that is common across instruments. Confidentiality of course must be maintained, so numbers but *no* participant names should be on forms and data files for identification purposes. Contact dates should be included on each questionnaire to distinguish that particular round of data collection across participants.
- ☐ There should be a link between the questionnaire ID variables and the tracking system (see Appendix C for a sample tracking log), especially links to information regarding household selection or other sampling criteria.
- ☐ There should be variables to indicate the intermediate and final status of the participant -- these should be linked to the tracking system.
- ☐ Variable names should be limited to 8 characters. (Some software allows 10, but if the need arises to transfer the data set to another type of software, having 10-character variable names can make the transfer much more difficult.) (See Appendix A.)
- ☐ As a rule, mnemonic variable names are easier for the analyst to use. FPMETHOD, PREGNANT, PARITY, etc. are easier to remember than VAR1, VAR2, VAR3. (See Appendix A.)
- ☐ The convention for coding yes/no questions is to use 1 and 0. However, some software packages change missing data to 0 so be careful! (See also Appendix A for more on yes/no questions.)

Codebook



Q. What is a codebook?

A. The codebook is a crucial piece of study documentation. The codebook should include:

- ☐ variable names
- ☐ field locations
- ☐ valid ranges for data
- ☐ explanation of codes for categorical data, including codes for "other/specify" fields that may be created after the form has been processed
- ☐ frequencies for selected variables
- ☐ use of any "special missing" values (for example, "not applicable" or "don't know")

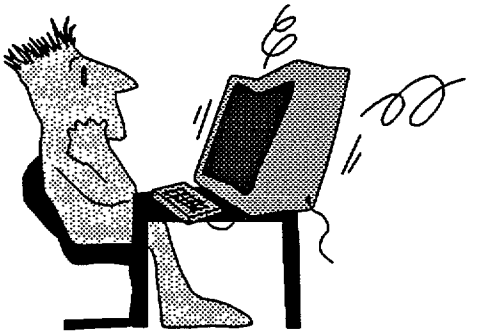
Note: A copy of each final questionnaires with the variable names beside each item should be part of the codebook.

Data Management

Q. What are the usual steps in data management?

A. The management of the data files involves a number of steps that will assure data quality. Remember the maxim: "Garbage in, garbage out." If you cannot rely on the accuracy of the data, the analyses are worthless.

Ideally, as many of the steps in data management as possible should take place as close to data collection (in time and space) as possible.



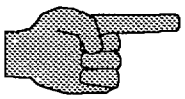
Data Management Steps

- Data Editing
- Data Entry
- Data Audits
- Data Checking
- Data Queries
- Data Changes

Editing

Editing should be done as soon as possible after the data are collected, so that the interviewer or focus group leader can answer questions and a follow-up visit can be made if necessary. The supervisor of data collection should check each form prior to data entry. This involves ensuring that there is no missing information, the handwriting is legible, the questions appear to have been interpreted correctly by the data collectors, the dates are correct, and the responses can be appropriately coded. If open-ended questions need to be coded, this is often done at, or just prior to, this step.

Keying/Data Entry



There are two basic modes for entering data into an electronic database (see Figure 1). In the "batch processing" mode, batches of data forms are sent to a data entry person, who types the data into an ASCII file (typically numbers in specific columns). Definition of the data base is carried out by the data analyst using statistical software, and the data are read into to database using "batch input." Data checking is then performed.

The second approach is "interactive processing." There are packages available now (e.g., Epi-Info) that allow "interactive processing." In this mode, a software program can be created that will prevent an invalid code from being entered and skip fields appropriately according to the responses already given. Sometimes consistency checks can also be programmed. The program shows a screen for data entry resembling the data form. Also, if the computer is at the scene of data collection, questions about the data may be brought to the attention of the field supervisor while the original data forms are in hand, reducing the necessity of going to the study files to resolve data queries. By these means, errors occurring during data entry are greatly reduced.

It is common for data under either data entry mode to be "double-keyed," that is, entered twice by two different people. This helps reduce on errors due to mis-entered data. Software is available that either warns the second typist that something was entered that does not match the first typist's entry, or compares the two files and notes discrepancies. Discrepancies should be resolved by a third party, if possible.

Data Audits

During data entry/loading, a proportion of original questionnaires should be checked against the data file and data forms to verify the accuracy of data entry. For small studies, 100 percent checking is possible, but more commonly, 10 to 20 percent are audited. If data are double-entered, there is less need for extensive checking against the data forms, but at least 10 percent of forms should be compared with data entries.

Data Checking

Using either mode of data entry, the database will need to be checked after the data are entered. If data were entered using batch processing, this checking will need to be more extensive. These checks are necessary:

Record level

- Checks for missing records and duplicate records. Decision rules should be in place for true duplicate records.
- Checks for incorrect ID numbers.

Item level

Many of these can be coded into data entry programs so inconsistent (and out of range) entries are immediately found.

- Consistency checks -- to ensure that responses make sense and are consistent. (See Appendix B.)
- Range checks -- to ensure that all codes entered are valid codes for that question. (If the data were entered interactively, and the data entry program was written correctly, there should be no invalid codes.)

Data Queries

There should be a procedure in place for handling questions regarding the data. Queries should be written and answers provided in writing and initialed by the principal investigator. Often there is a special "query form." A memo detailing how to handle commonly occurring questions should be disseminated to all concerned staff.

Query form contents:

- Date the discrepancy was noted
- ID variables for the form
- Incorrect or suspicious value
- Any notes regarding the question about this value
- Correct value
- Notes or explanations regarding the change of data (or explanation for why it was correct), and
- Signature or initials of the principal investigator.

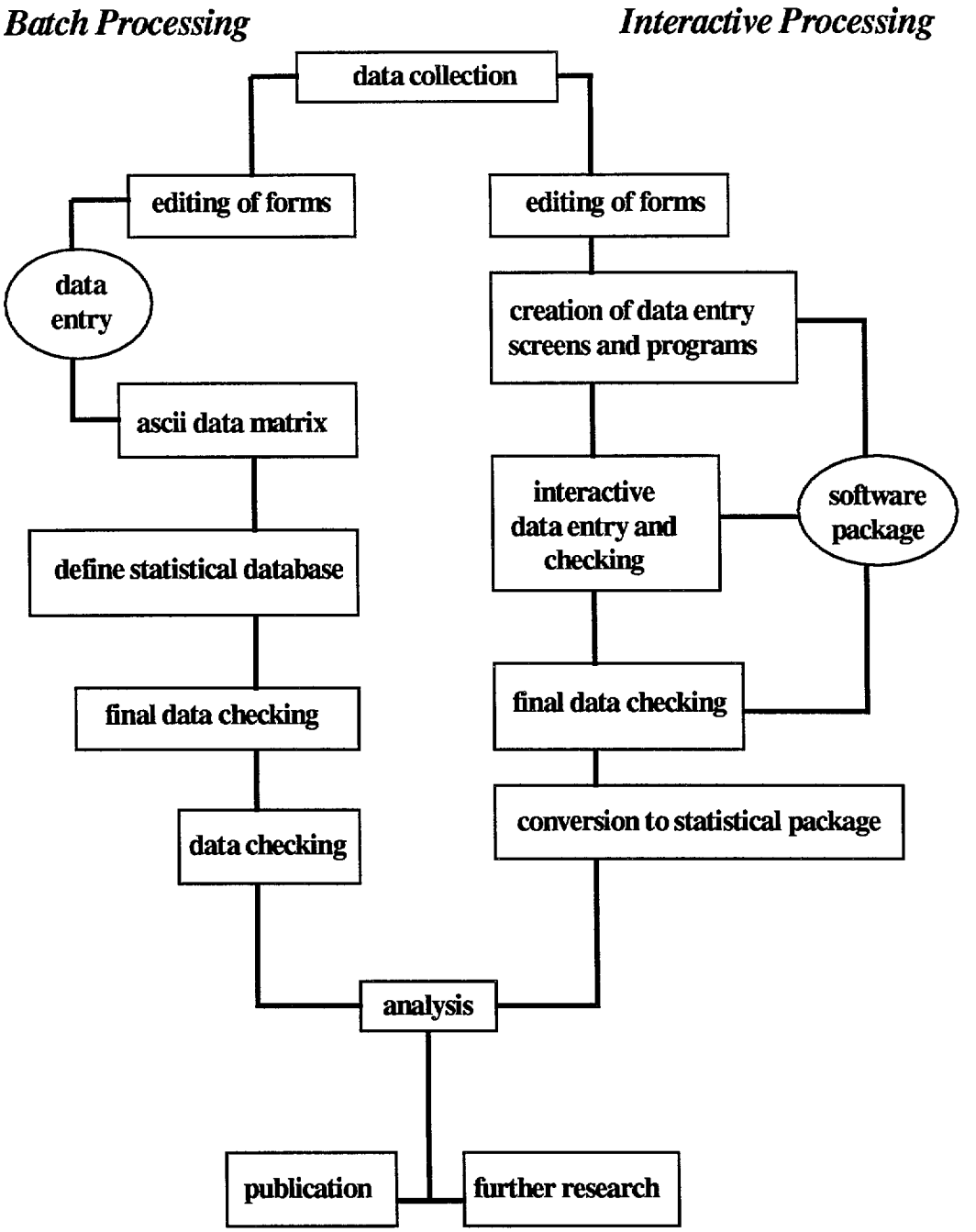
Data Changes

All changes made to the data already keyed should be documented. This documentation can be kept in a logbook or on query forms. Anything changed on the data file should also be changed (with an initial and a date) on the original data forms. The data in the data set must match that in the original data forms. If they do not, it will be very difficult to know which value is correct should a question arises about the data at a later time.

A copy of the original, unmodified data set must be kept. If changes are to be made to the data set, a new copy of it should be created first, and all changes made to this new copy.

Considerable time and energy can be spent on making changes to the data set. It is important to decide on a point at which the data set will be "frozen"; that is, no additional changes will be made after that time unless very serious discrepancies are found during data analysis.

Figure 1. Sequence of Events for Two Methods of Conputerized Data Collection

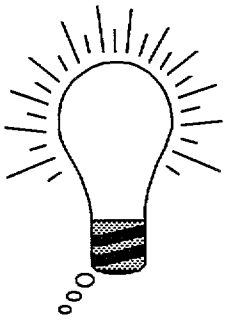


Adapted from Romensz, N. (ed.) (1989). *Data Management and Clinical Trials: EORTC Study Group on Data Management.* New York: Elsevier: 100.

E. Data Analysis

Q. What is an analysis plan?

A. An analysis plan outlines the steps that will be followed to analyze the data, the variables of interest -- including their operational definitions, the statistical tests used, and the subgroups of interest. The analysis plan should be described both in the subagreement and study operations manual, and should be reviewed before beginning data analysis. It will be helpful to include a copy of the original plan with the report of the analysis. The analysis plan may be revised prior to or during the actual data analysis. Any revisions should be noted in the analysis plan. Table shells (or dummy tables) for presenting the research findings are often included in the analysis plan, and help in the design of the questionnaire (avoiding extraneous questions) as well as in focusing the analysis plan itself. It is highly recommended that the table shells be completed prior to the start of data analysis, since table shells are enormously helpful in organizing the data analysis.



Q. What is an analysis data set?

A. Modification to the original data usually will need to be made before the appropriate analyses can be conducted. The new data set can therefore be called the "analysis" data set, and all modifications must be documented. A printout of the program statements from the statistical software used will suffice. Useful information to be included:

- ☐ Justification for the exclusion of any subjects or data forms from the analysis
- ☐ Handling of missing data
- ☐ Recodes, creation of new variables
- ☐ Merging of data sets

It is imperative that one or more copies of the original, unmodified data set be kept, and that all changes and analyses be carried out on a separate copy of the data set. Then, should any unforeseeable problems occur in the analysis data set, the researcher always has recourse to the raw data.

Q. How is an analysis data set documented?

A. The easiest way to document the data analysis process is to print out the .LOG file and keep it attached to the analysis results. Some software (e.g., SPSS) prints the log statements (program statements) along with the results. Other packages (e.g., SAS) prints the output to a .LIS file and the program statements to a .LOG file. Having the program statements with the results is the only way of knowing for certain what was actually done.

It is useful to add comments to all data programs: for example, who the programmer is, when the program was first created, when it was last modified, what its purpose is and any important information about how the data are being handled. (Some software packages will include some of this information automatically.) Comments can be included throughout the program to explain the program statements.

While people have different styles of performing and documenting data analyses, it is usually best to limit the number of steps from the first data file creation to the analysis output. That is, create as few new data sets as possible.

For example:

- ☐ Create one program to load and label the variables, assign missing values, perform recodes that will be part of every analysis and exclude cases that will be permanently excluded.
- ☐ Write another program to recode and manipulate the data in the manner required for the current analysis.
- ☐ Run the appropriate statistical tests. The statistical analysis program may be separate and can be prepared for groups of table shells.

If there are many steps (separate programs) to the final analysis data set, and if a change has to be made to the original data, all those steps will need to be retraced. Retracing all the steps is time-consuming and introduces opportunities for error since any given step has a chance to be omitted. It is also harder to determine what was done in the analysis (i.e., how a particular variable was defined, how some missing values were handled) if the documentation is in a number of different programs. Additionally, there may be confusion over which is the most current or the "official" data set if several exist.

Titles and footnotes (including job name and date, program that was used) should be included in the output whenever possible. Appropriately titled and labeled output tables are much easier to interpret and use.

The data analysis set prepared by the in-country researchers will be the official data set for the study. It is very important that the documentation of this data set be clear, thorough and complete.

Section II

Data Analysis Verification

Q. What is data analysis verification?

A. Data analysis verification means that an *independent* check is carried out on all of a study's analyses and reports to ensure that statistical analyses have been performed correctly and that results are reported accurately. When possible, two different analysts should run the statistical data analyses to ensure independent checks.

For the sake of brevity, the term "verification" will be used to refer to data analysis verification in the following sections.

Q. When is verification conducted?

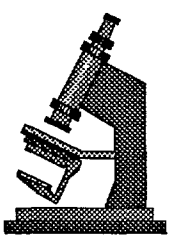
A. Verification typically is carried out after results of statistical analyses are written in a draft project report or draft journal articles. If no major flaws are identified, both the correctness of the data analysis and the accuracy in reporting can be attested to at the same time. However, if major errors are identified, the statistical analyses will have to be rerun and results will have to be rewritten and rechecked.

To avoid the necessity of rewriting, verification may be done as soon as the major statistical analyses are completed, before results are drafted in reports or articles. Depending on the complexity of the analyses and the lag time between analysis and report writing, this approach allows researchers to check repeatedly the correctness of the statistical procedures prior to report writing.

For some research projects, verification may take place *both before and after* results are written, depending on: 1) personnel and time resources available; 2) complexity and number of statistical analyses and 3) lag time between statistical analyses and reporting.

Q. Whose responsibility is the verification process?

A. The principal investigator is responsible for implementing the data verification process according to the guidelines set forth in this document. However, under some circumstances, parts of the verification may also be carried out at FHI. In-country researchers and FHI technical monitors should plan for data verification as soon as possible. Ideally, the planning for data verification begins



when table shells and analysis plans for the research study are drafted.

The verification process, as well as the documents needed to facilitate the process, should be part of each project's monitoring plan. FHI technical monitors and in-country researchers are advised to include as part of the agreed monitoring plan the initiation, organizational set-up, scope, timing and technical assistance for data verification.

Q. How does a researcher prepare for verification?

A. Some of the key data verification issues that country researchers and technical monitors will address during the planning stage include:

- Specify **which analyses require verification.**

Ideally, all statistical tables and figures in the project report should be verified. However, depending on the complexity and type of analyses and available personnel resources, the in-country researchers and FHI technical monitors may need to prioritize statistical tables and analyses to verify.

The scope of the data verification (and in some instances, the type of sophisticated data analyses) to be completed in the country may be affected by the availability of computing and personnel resources. It is very important to specify early in project implementation whether all or part of planned data verification will be done at the in-country research organization or elsewhere, since the written documentation needed for an outside data verification process is more detailed than when verification is undertaken within the same research organization. Moreover, verification outside of the research organization will require more resources and additional lead time between the original and verified analysis.

- Organization of the **data verification staff.**

As early as possible, the PI should designate the project staff who will conduct the original data analyses and those who will conduct the data analysis verification. Depending on the scope of data analyses, the original data analysts and verifiers may consist of a statistician, computer programmer and subject matter specialist.

The programmer/analyst who conducts the verification should be different from the one who prepares the original statistical tables and conducts data analysis. In some projects where this is not possible (for example, if there is only one statistical programmer or analyst), it is recommended that the original analysis and verification be done with some lag time. The extent of the lag time will depend on the overall project duration. This lag time allows the analyst to approach the analyses and reports with "fresh" eyes.

In instances where the software appropriate for more advanced analyses may not be available in the research institution or in the county, the in-country researchers and FHI technical monitors should agree on the scope of data analysis and verification to be carried out either at FHI or at a consulting institution within the country.

- Specification of the **documents needed** for data verification.

The following data sets and programs will be needed by the data verifier:

- ☐ Relevant data sets including **original, modified and analysis data sets** with their location and file structure.
- ☐ **Computer programs** that created the above data sets. Each data set should be linked with a specific program.
- ☐ Log, anecdotal information or other **documentation explaining data handling** (e.g., the handling of missing values) In general, the handling of missing data should be described in the data management manual. More details on data management procedures are in Section I.
- ☐ **Format programs** used, if any.
- ☐ Sample copies of relevant **data collection forms**. Variable names should be on the forms or codebooks provided. A codebook and a copy of the questionnaire(s) with variable names and additional codes (if necessary) written alongside the items are very helpful to the data analyst. Tips on naming conventions are given in Section I and in the Appendixes.
- ☐ A **list of created variables** (not on data collection forms) and their definitions. Documentation can be in the program listing or in the analysis plan. The specifications for created variables must be in the analysis plan. When data are analyzed using complex sample survey procedures, the sampling design and the programs used to calculate weights should be given to the data verifier. (See other details on sampling design and analysis weights in Section I.)
- ☐ **List of computer programs that generated output for tables.** Each table should be linked with a specific program.
- ☐ **List of computer programs that generated outputs for findings cited in the paper but not in tables.** Linkage with specific programs should be noted.
- ☐ An **annotated copy of the manuscript** (if already written), including names of programs used in the analysis..

- Steps in the **actual conduct of data verification.**

The steps in the actual data verification are described in the coming pages. The in-country researchers and FHI technical monitors will need to decide which of the steps in the data verification process are appropriate for their projects.

- **Documentation of the verification process.**

Results of the planning, implementation and evaluation of the verification process can be incorporated in the monitoring plan for each project. See steps in data verification below for more details.

- **Resources** including manpower, computing time, and the schedule and timing of data verification.

In planning for verification, the in-country researchers and FHI technical monitor should decide on the extent of manpower to be used as well as the schedule for and timing of the data verification process. Such resources will be reviewed occasionally to meet any needs that may arise during the original data management and analyses.

- **Technical assistance** needed.

The in-country researchers and technical monitors need to discuss whether technical assistance is needed, the nature of any technical assistance to be provided, and when such assistance would be needed during the data verification process.

As specified in the project sub-agreements, the FHI technical monitors or FHI consultants will provide technical assistance to the researchers in the various phases of the research projects. As it applies to data verification, such technical assistance may include:

- ☐ Direction on types of statistical analyses appropriate to the data and research questions addressed in the study;
- ☐ Provision of sample survey software like EPI-INFO for projects where respondents were selected using complex sampling procedures;
- ☐ Planning and clarification of the data analysis verification procedures.

Q. Are there any guidelines in the data verification process?

A. To complete the data verification process, the data verifier does the following:

- ☐ Reviews formation of data sets and creation of new variables. Recreates analysis variables including outcome, control and independent variables. Checks formats of raw and recoded variables against codes in data collection forms. Those variables that need recreation must be identified during the planning stage.

- ☐ Reviews the stated analysis plan. Verifies that it was followed. Notes any deviations from the analysis plan.

- ☐ Lists statistical tests used. Comments on their appropriateness, including fit of the data with the assumptions of the completed statistical tests:

Assumptions to Check

- Are sample sizes, especially for analysis involving subgroups rather than the full sample, sufficiently large for adequate power for the statistical test?

- Frequency distributions of analysis variables, especially the outcome variables must be checked for shape, skewness and outliers (assumptions of normal distribution and homoscedasticity). If the error terms in fitting models to data are available, a plot of the error terms must also be checked for shape, skewness and outliers.

- Create scatter plots of outcome versus key explanatory variables. The scatter plot generally reveals the pattern of association, if any exists, between two variables.

- Analysis weights are used in the analysis of data from complex sample survey. If analysis weights are not used, indicate the reasons why. Depending on the agreement between the principal investigator and FHI technical monitor, analysis weights may also be subject to data verification.

- ☐ Uses independent computer programs to reproduce all statistical tables and analyses to be verified. Notes any discrepancies between the original and verification runs. In instances where there is only one programmer, it is very important that all created variables be crosstabulated with the variables used for creation and that someone other than the programmer check the accuracy of variable recoding. As mentioned earlier, if there is only one data analyst, it is important that the original and verification programming be

done with sufficient lag time. This allows an independent programming approach to be developed to check the correctness of the first approach.

- ☐ Checks numbers in text against those in tables. Notes any discrepancies. Checks whether the denominator(s) used in reporting percent distribution is (are) the most appropriate to use in answering the study objectives. Checks footnotes and explanatory text to ensure they accurately reflect how analyses were conducted.

- ☐ Reproduces numbers in report section of text that are not found in statistical tables.

- ☐ Confirms p-values of all tests reported in the text or tables. In instances where p-values are not reported, notes the reasons why.

- ☐ Comments on significance levels and consistency in reporting p-values. Notes if multiple testing or selection bias is a problem.

- ☐ Verifies numbers in discussion section of the text. Comments on any numbers that are not supported by the analyses conducted.

- ☐ Confirms whether statements made in the text are supported by the results of the statistical analyses. Pays particular attention to "implied" results.

- ☐ Maintains continuing communication between the project investigator and the original data analyst/programmer. As soon as the verification process is started, it is important that ongoing communication exist between the original data analyst/programmer and the data verifier. The data verifier must communicate questions about the verification process with the principal investigator and original data analyst/programmer. Discrepancies in the original statistical tables and verified tables must be resolved with the original data analyst/programmer. If there are major discrepancies between the two sets of tables, a third party may be needed. This may include the involvement of the project director if the discrepancy involves a data inconsistency or a difference in the interpretation of findings.

- ☐ Prepares report and documentation of the verification process. The documentation of the verification process will list the variables recreated, the computer programs used to independently rerun the statistical analysis and the new data sets used for data verification. The software used and version of software used must also be specified in the report.

- ☐ Archives the data sets and programs used in the data verification process. If the data sets used for verification are converted from one software package to another, the original and converted data sets should be archived. The log of the conversion should also be archived.

Verification Results

- Original analyses are adequate and no revisions are needed.
- Original analyses require minor revisions.
- Original analyses require major revisions calling for a second set of verification.
- Verification could not be completed, including the reasons why.

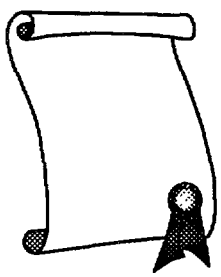
Study Closeout Procedures

Certain items need to be carefully labelled and stored locally in a secure place (safe from humidity, insects and other hazards) before the end of the study. These items include:

- ☐ The analysis plan for all types of data collection (submitted prior to implementation).
- ☐ Local language (plus English) versions of final, pretested data collection instruments for quantitative components of the study.
- ☐ Computer printouts of frequencies, cross-tabulations and additional statistics required for data analysis.
- ☐ A final report (both paper copy and on diskette) in the local language as well as in English.
- ☐ Final diskettes with data from the study, including codebook.
- ☐ Verbatim transcripts of all in-depth interviews (in local language and in English, if feasible).
- ☐ Any papers prepared based on the study.

The PI and FHI monitor should discuss which of these should be sent to FHI and at what point in the study.

Although study documentation is a demanding activity, if it is not done well, the researcher ends up with many headaches at the close of the project. Good documentation leads to smooth verification and a more enjoyable research experience for everyone involved.



Glossary of Terms

Analysis data set: the "clean" version of the study's data set that will be used in the data analysis.

ASCII: A generic computer file format that can be readily converted to any other program format.

Batch input: The conversion of data from an ASCII file to a database file using a computer program. This allows groups, or batches, of data to be processed at the same time.

Batch processing: A method of data entry in which data forms are put into an ASCII file in groups, then the data are converted as a group ("batch input") into a database form by a computer. The data are not checked until later.

Consistency checks: A type of data check that involves verifying the logic of responses to items on a data collection instrument. Can be done both during and after data editing.

Data audits: The process of checking a computer data file against the form(s) from which the data come.

Data checking: A number of methods that ensure the accuracy of the data in a data set.

Data collection form: An individual copy of a data collection instrument or questionnaire.

Data collection instrument: The method used to gather information in a study. Examples of data collection instruments include questionnaires and focus group guidelines.

Data editing: The process of checking for and correcting data errors. Data editing can also refer to a step in the process when completed forms are reviewed and coded for the first time.

Data entry: Also called "keying" or "keypunching." The process used to transfer data from a data collection form onto a computerized database, usually by typing responses into a program. .

Data queries: Questions regarding data values.

Data verification: The process of ensuring the quality and accuracy of the analyses used in a report or paper.

Documentation: The keeping of a written record of each step in the research process.

Double keying: A method of data quality assurance whereby two separate individuals enter the same data from a data form. The two versions can then be compared, and discrepancies identified.

Duplicate records: The result of double keying. "Duplicate records" is also used to mean an

error in forms completion (resulting in two records for the same person).

Field location: The placement of an item on a page or electronic screen.

Field operations manual: The part of the study operations manual specifying how data will be collected.

Interactive processing: A method of data entry that accepts only valid codes to be entered; sometimes includes skip patterns based on data already entered from a given form.

Item: A question on a data collection instrument.

Keying: See data entry.

Programs: A set of instructions for a computer or software package to follow.

Query form: A form used to handle discrepancies in the data in a data set; usually includes the date the discrepancy was noted, the ID variables for the form in question, the value in question and any notes pertaining to the data, as well as an area for the principal investigator to respond to the query.

Range checks: A type of data check that involves verifying whether a response to an item is valid. For example, the valid codes for an item are 0 = no, 1 = yes, 2 = no answer/other, a range check would flag any code 3's as invalid. Range checks can also detect suspicious but not necessarily erroneous values, which can then be confirmed (for example a value of 43 kg for the variable weight).

Study operations manual: A written research plan detailing each step of the study.

Tracking system: Included in the field operations manual, the tracking system documents how the sampling plan was implemented.

Verification: See Data verification.

APPENDIX A

Naming Variables: What's in a Name?

The name of a variable is very important in data entry and analysis. Names that are consistent and informative can save time at every level of data entry, management and analysis. Unfortunately, no name scheme is perfect!

1. Keep variable names to 8 characters or fewer.

Epi Info allows you to create variables that exceed 8 characters. However, many software packages limit variable names to 8 characters. If you create a variable with more than 8 characters, you may have a problem should you try to switch to another program: If there are two variables whose names are the same for the first 8 characters, their names may become identical in the new program if you use Epi Info to convert the variables. For example, "PREGNANCY1" and "PREGNANCY2" would both be converted to "PREGNANC." One way to prevent this is to use the program DBMSCOPY to convert from one database to another. DBMSCOPY will automatically rename the offending variables when it converts the Epi Info data into a data set for another software package. However, one is left with variables names that were unplanned. Renaming and/or working with the unplanned variable names takes more time.

2. Use names that describe as accurately as possible the variables they represent.

In creating a variable name for the question "Are you satisfied with your method of family planning?," you would want the name to focus on satisfaction and family planning method. One possible name might be "SATMETH." If you were to simply call it "VAR10" because it was the 10th item on the questionnaire, you would not be able to identify the variable by its name. The names "METHOD" and "SATISFY" are reasonable choices, but they are not as informative about the question (satisfaction with the method). Also, if there are multiple items about family planning methods or satisfaction with various things, the names "METHOD" and "SATISFY" could be troublesome. Avoid using "OTHER" to refer to an "other/specify" field. Usually there will be many of these in a given study.

3. If you must shorten a word in naming a variable, make it end in a consonant.

Names that follow this convention are most pleasing to the eye and easier to decipher. For example, "years of education" could be coded as "EDUCATIO" or "EDUC." The first choice, "EDUC" is easiest to understand. Sometimes less is more.

4. Give variables unique names on each form when using multiple forms.

If there is more than one form per participant, you must assume that you will be merging data across forms. It is important that every form have distinct date variable names. For example, if an initial questionnaire asked about prior contraceptive use and a follow-up questionnaire asked about current use, you would not want to name both variables "FPMETH." Typically, if these data sets were merged, the value for one of the variables would be overwritten by the other. The variables can be renamed during analysis, but it takes extra effort to do so, it can be confusing and it introduces more opportunities for error. An easy way to avoid this is to let the first few characters denote the form name. For example, "BLFPMETH" would be the name for the baseline (BL) questionnaire item (for family planning method), whereas "FUFPMETH" would be the name of the follow-up questionnaire item. Another option would be to limit variable names to seven letters and to add a letter at the end that indicates which form is being used.

5. Be consistent when coding "yes/no" variables.

It gets confusing if the codes for some variables are 0=no/1=yes and for others are 1=no/2=yes, or even 0=yes/1=no. The person analyzing the data must be very careful about which variable he or she is examining.

Reverse coding generally is used to reduce response pattern bias in opinion/attitude items. Responses to such items must be handled very carefully, especially if the data are to be combined with data from items that were not reverse coded. Often new, derived variables are created from the reverse-coded variables and these new variables are then used in the analysis data set. In this way all variables share the same coding scheme. It is recommended when recoding that the higher values be used to denote "more" of the dimension of interest. For example, if a *higher* score on an item means the respondent is more introverted rather than extroverted, the variable would be called "INTROV" rather than "EXTROV."

6. Handle other coding with care, too.

When it comes time to analyze data, researchers who coded well will be rewarded by knowing what various codes truly mean, and by being able to differentiate between sometimes ambiguous responses. A little foresight can help overcome the obstacle of answers that might be coded as "0," for instance: A "don't know" response, "not applicable" or the lack of a response where there should have been a one all could possibly be coded as "0" in the data set. Valuable information may be lost in the analysis if no differentiation is made among these responses, since they mean different things. Separate codes for each type of response is highly recommended.

APPENDIX B

Creating Consistency Checks: Can Men Really Get Pregnant?

It is important to check the consistency of data across forms as well as on the same form. In addition to checking the range of each response, look for inconsistencies in responses. A specific answer to one question may fall within the right range but be inconsistent with the answer to another question. Assume that people will answer inconsistently when given the opportunity! Tips for checking data:

Across-Form Checks.

Check the linkage.

A common mistake made when designing across-form consistency checks is to assume that the forms are correctly linked (that is, that all forms with the ID 0052 belong to the same person) and to check only those items that would represent respondent errors. No matter how good the data collection methods, there is always the chance that a typographical or other error will result in an improper linkage. This may be particularly true for studies done in "natural" settings -- for example, in a busy clinic where many people are being processed at the same time. When designing across-form consistency checks, assume the possibility of incorrect linkage.

If the nature of a study is particularly prone to this type of problem, the researcher may want to incorporate certain items into the forms (for example, date of birth on a follow-up form) to confirm linkage.

Think of other possible inconsistencies.

If you have interviewed all women of reproductive age in each household, the total number of interview forms for the household should be less than or equal to the number of individuals in the household and equal to the number of women of reproductive age (if this information is available).

Within-Form Checks.

Be attentive to skip patterns.

People do not always follow skip patterns.

Example 1: A reproductive history questionnaire used for both men and women would have men skip patterns about pregnancy. "How many times have you given birth?" would be answered by women only. In checking for consistency, you'd want to make sure the question about giving

birth was missing data for all men.

Example 2: "Have you ever had an STD? If yes, please specify." Most participants will answer correctly. However, some will likely (a) say "yes" and not specify, or (b) say "no" and specify.

Look for missing data.

Sometimes data that are critical to the study's objectives will be missing. You might want to check and print out any data that you consider absolutely necessary to collect.

Think of other possible inconsistencies.

It is unlikely that a 20-year-old woman has 10 children. However, there may be someone in the study who has 10 children by age 20. So while some values are truly inconsistent, some responses are rare but possible.

Focus on the inconsistencies involving the most important variables.

Consider the cost in time it takes to create consistency checks versus the gain in data accuracy.

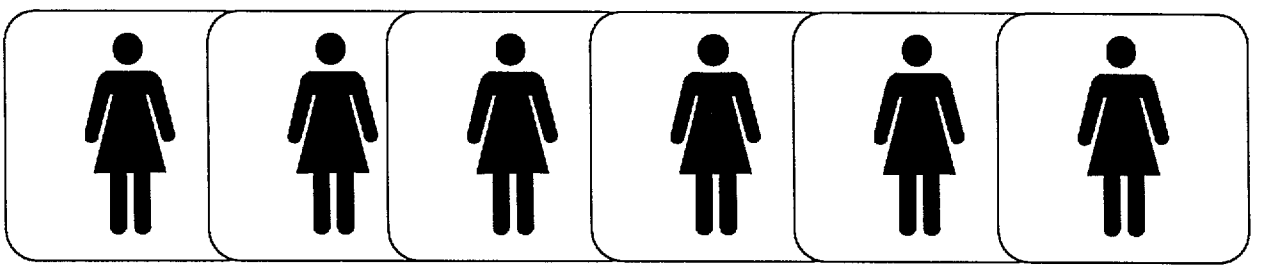
APPENDIX C

Example of an Interviewer Log Sheet

YOUTH STUDY
1995
Interviewer log sheet
Geographical area:_____
Household address:_____
Interviewer name:_____

Results Codes
1: Completed interview
2: Refused interview
3: Eligible not at home
4: No one eligible
5: Vacant/No one home
6: Other/Specify

House No.	Eligible Men 15-19	Eligible Women 15-19	Eligible Women 20-24	Eligible Chosen	Visit1 Date	Visit1 Result	Visit2 Date	Visit2 Result	Final Visit Date	Final Visit Result	Notes



APPENDIX D

An Example of a Sampling Plan

The following is an outline of a sampling plan for the survey of family planning knowledge among adolescents in Dakar, Senegal.

Target Population

Married and unmarried females aged 15 to 24 and males aged 15 to 19 living in housing units in Dakar and Pikine districts in Dakar City, Senegal. As of 1989, the study population numbered 209,044, with 52.4 percent residing in Dakar and 47.6 percent residing in Pikine districts.

Survey objectives and data-based objectives

The survey objectives are:

- To assess the knowledge, attitudes and practices of adolescents regarding family planning.
- To identify actual and perceived barriers to family planning services.

The data-based objective is:

- To provide estimates of characteristics that occur with a frequency of at least 10 percent in the population to within plus or minus 30 percent of the true proportion, with a 95 percent probability.

More specifically, the data-based objectives are to estimate the proportion of females aged 15 to 24 who married following an out-of-wedlock pregnancy and to estimate the proportion of males aged 15 to 19 who are knowledgeable about key family planning practices. (These objectives were quantified by providing estimates of the proportion of females with out-of-wedlock pregnancy at 11 percent and proportion of males with FP knowledge to be at 50 percent.) These objectives are derived by identifying the most important variables that the researchers would like to estimate with reasonable precision. The precision of the estimates are stated in terms of the differences from the true population values that the researchers find acceptable with certain probability. This is where the 95 percent probability, also called a 95 percent confidence level, comes in.

Sample Size

A total of 1,024 men aged 15 to 19 will be recruited into the study, of whom 75 percent are expected to complete the interview. A total of 1,566 women aged 15 to 24 years will be recruited, of whom 90 percent are expected to complete the interviews. The female sample sizes are sufficient to yield estimates of proportion of women who got married following pregnancy

to be within plus or minus 30 percent of the estimated true value of 11 percent with 95 percent probability. It is assumed in the calculation of the sample sizes that the design is 2, doubling the estimates of variance relative to simple random samples of the same size.

Selection procedures and sampling frame

The observational units, females aged 15 to 24 and males aged 15 to 19 are selected using a stratified clustered household survey. The first stage sampling frame consists of a list of districts in Pikine/Dakar areas. Associated with the list of districts are the total number of housing units, total males and females and total population in each district. Districts within the same area are grouped and classified according to their total population. Districts within Dakar and Pikine are classified as large, medium or small, for a total of six strata. Large and medium districts are to be subdivided so that the average size of a subdistrict is close to the average size of the small districts. With 25 percent of housing units expected to have an eligible person, all housing units in sample districts/subdistricts will be screened for the presence of eligible household members. With an average size of 80 to 100 households per district or subdistricts, 20 to 25 housing units are expected to contain an eligible sample person. If more than one eligible person is in the household, the interviewer will select a sample person from each sex-age group using a sample selection table. Instructions on using the selection tables are given in the study manual. The designated sample person will subsequently be asked to undergo a personal interview.

The total of persons sampled is proportionately allocated to the sampling strata. The total number of districts to be selected is determined by the total number of sample persons expected from each stratum.

(Include here a table of the population counts by strata, number of eligible persons to be selected from the strata and the number of districts to be selected should be included. From such a table, the probability of a given person being selected can be calculated. The probability of selecting a person is a derived from the probability of selecting the district, the probability of selecting the household and the probability of selecting the sample person from the list of age-sex eligibles in the sample household.)

Estimation procedures

- Separate analysis weights will be constructed for the female and male sample.
- Sampling weights, which are the inverse of the probabilities of selection, will be adjusted if response rates differ significantly across strata and levels of key variables of interest differ across strata.
- With clustering and stages of sampling, complex sample survey software will be used for analysis especially where statistical tests of hypotheses will be done. Design effects will be estimated. Design effect is the increase in variance of an estimate from a sample selected using complex sampling procedures relative to the variance of estimate made from a simple random sample of the same size.

Elements for tracking system

The following items should be included in the study tracking system to ensure that appropriate weights are included:

- sampling variables (strata/psu/housing unit)
- number and dates of visits (maximum has to be determined by field staff and supervisors) for screening
- result codes for each visit
- final screening codes
- for those households successfully screened, presence of eligible sample persons
- indicator for selection of eligible sample persons
- number and dates of visits for interview of sample person
- result codes for interview at each visit of sample persons
- final person interview result code
- status of interview questionnaire:
  -- edited by field supervisor/date
  -- sent for keying/dates
  -- keyed/double keyed/dates
  -- sampled for audit
  -- results of audit
- decision rules for completion rates; when to accept refusal/number of fall backs and final disposition of cases

Checking of data elements may render a completed interview to be considered ineligible, or with incomplete data.