FINDING AND UNDERSTANDING VARIABLES IN DHS DATASETS

Before analyzing DHS data, it is important to understand DHS data structure and know the different types of DHS data files. This job aid provides a brief overview of the standard DHS data files and an introduction to the data structure.

DHS data files are large and complex, with thousands of variables and tens of thousands of cases. It's not always easy to find the information or variables you're looking for or immediately understand the contents of a DHS data file. This job aid explains:

- how data are collected in a DHS questionnaire is transformed into a DHS Standard Recode data file;
- how to use the recode manual to understand exactly what is in each variable; and
- how to find the variables you will need in your analysis.

DHS standard recode files

Types of DHS questionnaires

DHS recode manual

Other (quick) ways to find your variables

Summary

USAID
FROM THE AMERICAN PEOPLE

DHS STANDARD RECODE FILES

Since 1984, DHS has implemented a Standard Recode, which means that certain pieces of information collected in the questionnaire – for instance, whether a woman lives in an urban or a rural area – will always be recorded with the same variable name in every dataset.

Variable names/numbers and definitions are standard in **every** survey in **every** country and sometimes also between files:

- **v025** is ALWAYS urban/rural residence in an IR file
- **hv025** is ALWAYS urban/rural residence in an HR file

However, the tradeoff for this convenience is that the variable numbers in the dataset do not match up with the question numbers in the questionnaire.

This means that we could write a program referring to **v025**, and the program would run perfectly on every single dataset. For example, we could use the same program on every Tanzania survey.

Why don't the variable numbers match the questionnaire numbers in recode data file?

The organization of the recode file is based on the core questionnaire, *but:*

- The core questionnaire has changed over time
- Some countries add additional questions
- Some variables aren't in the questionnaire, but are constructed and included in the recode file (e.g., wealth quintile, unmet need)

Instead of making each recode variable match the questionnaire numbering, DHS uses a standard numbering for each variable, which varies from survey to survey.

- Enhance reproducibility: A Stata do file or SPSS syntax file that you write for one survey can be used on any other survey because the variable names are the same with the exception of survey specific variables.

TYPES OF DHS QUESTIONNAIRES

- All DHS surveys use a household questionnaire and an individual woman's questionnaire. Most surveys also include a man's questionnaire. Biomarker questions used to be included in the household questionnaire; starting with DHS 7, there is a separate biomarker questionnaire.
- **Standard content** of DHS surveys generally include (DHS 8 core):

HOUSEHOLD QUESTIONNAIRE

1. Basic characteristics of **all household** members (relationship to the head, age, sex, education, etc.)
2. Household characteristics (assets, water sources, sanitation facilities, ownership of mosquito nets, etc.)

WOMAN'S QUESTIONNAIRE

1. Respondent's background
2. Reproduction/birth history
3. Family planning
4. Pregnancy and postnatal care/maternity history
5. Child immunization
6. Child health and nutrition
7. Marriage and sexual activity
8. Fertility preferences
9. Husband's background and woman's work
10. HIV/AIDS-related knowledge and behaviors
11. Other health issues such as tobacco and alcohol use
12. Contraceptive calendar *(not included in every survey)*

MAN'S QUESTIONNAIRE

1. Respondent's background
2. Reproduction
3. Family planning
4. Marriage and sexual activity
5. Fertility preferences
6. Employment and gender roles
7. HIV/AIDS-related knowledge and behaviors
8. Other health issues

BIOMARKER QUESTIONNAIRE

1. Height, weight, and hemoglobin measurement for children under 5 years: Found in PR, BR, and KR files
2. Height, weight, hemoglobin measurement, and HIV testing for women 15-49: Found in PR and IR files
3. Height, weight, hemoglobin measurement, and HIV testing for men 15-49 or 15-54, depending on the survey: Found in PR and MR files

NOTE:

Used to be part of the household questionnaire. Biomakers are not standard and based on requests from the country.

TYPES OF DHS QUESTIONNAIRES: CONTENT OF RECODE FILES

CONTENT OF RECODE DATASET: IR, KR, and BR FILES

VARIABLE:	DESCRIPTION:
v000-v191	Respondent's background
v201-v244	Reproduction
b0-b20	Birth history
v301-v3a09b	Contraception
m1-m78j	Maternity history, including postnatal care
v401-v482c	Health (women) and child's and mother's nutrition
h0-h80g	Child's health
hw1-hw73	Child's biomarkers
v501-v541	Marriage and sexual activity
v602-v634	Fertility preferences
v701-v746	Husband's background and woman's work
v750-v791a	HIV/AIDS
v820-v858	Additional HIV/AIDS variables
v801-v815c	Interview characteristics
vcal/vcol	Contraceptive calendar data
s…	Country-specific variables

NOTE:

We can see that the variable names are generally in the same order as the standard questionnaire: the first variables (*v000-v191*) are about the respondent's background. The variables beginning with *v200* contain information about her reproduction, followed by variables that begin with *b*, indicating the birth history. Variables beginning with *v300* are about contraception, etc.

In Stata, all letters in variable names (*v, b*, etc.) are lower case and must be written as such. In SPSS, they are in upper case, but can be written in lower or upper case.

CONTENT OF RECODE DATASET: HR and PR FILES

VARIABLE:	DESCRIPTION:
hv000-hv044, hhID	Household basic data (cover page)
hv101-hv129, hvIDX	Household schedule
hv201-hv234	Household characteristics
ha…	Woman's height, weight, hemoglobin
hc…	Children's height, weight, hemoglobin
hb…	Man's height, weight, hemoglobin
hml…	Mosquito nets variables
sh…	Country-specific household variables

NOTE:

As with the woman's questionnaire, the household datasets have the information recorded in the same general order in the dataset as in the questionnaire.

In older surveys, height and weight of children were only collected for children of interviewed mothers, and so the data were recorded in the IR/KR/BR files, not the HR/PR files. Also, men's height/weight are not regularly collected in every survey.

The unit of analysis for the HR file is the household (each observation represents one household) and the unit of analysis for the PR file is a person in the household (each observation represents one person- the child, woman, or man).

Naming conventions for variables – general overview

VARIABLE NAME	MEANING
vxxx	Standard variable (**women**)
bxx	Birth history (**women**)
mxx	Maternal care history (**women**)
hxx	Health history for **children** (collected from women)
hwxx	Height and weight information for **children of interviewed women**
sxxx	Country-specific variables (**women**)
mvxxx	Same as V variables, but for **men**
smxxx	Country-specific variables for **men**
hvxxx	Variables for the **household file**
shxxx	Country-specific variables collected for **households**

COUNTRY-SPECIFIC VARIABLES, representing questions that are not included in the DHS core questionnaire, begin with an "s" followed by the survey specific question number.

The **MEN'S FILES** follow the same naming convention as the women's files, except that all variables that begin with a "v" in the women's files begin with an "mv" in the men's files, and variables that begin with an "s" in the women's files begin with "sm" in the men's files.

Variables in the **HR/PR FILES** begin with hv instead of v as in the IR/KR/BR files. Special variables will begin with sh in the HR/PR files and s in the IR/KR/PR files. Some of these variables may match for instance hv025 are v025 are both place of residence. However, this is not always the case and you need to check the variable description.

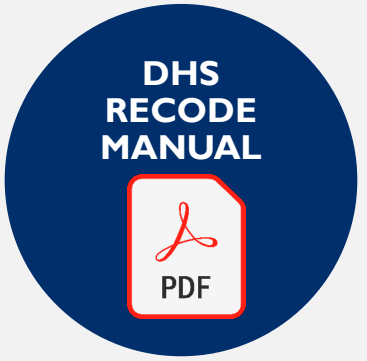
DHS RECODE MANUAL

- To access the manual, go to https://www.dhsprogram.com/publications/publication-dhsg4-dhs-questionnaires-and-manuals.cfm
- To help explain the contents of the Standard Recode files, the DHS created a recode manual, which is your guide to the recode data files.
- The recode manual provides the information necessary to understand recode datasets.
- The recode manual has two parts:
  - The first part is a general discussion of the recode file.
  - The second part provides a description of each variable in the data file, giving additional information that is not available in the datasets.
- Chapter 1 of the Guide to DHS Statistics also contains information about the recode file variables.

Why do I need a recode manual?

- To understand what the variables are and how they are defined, use the recode manual!
- Variable labels are short in the data files. Check the recode manual for a full description of the variables.

Click to open



More details about DHS variables

- **No decimal places in the recode variables**
  - For example, Body Mass Index values are stored in data files as 1850 rather than 18.50
  - Variables v437 (women's weight in kg), v438 (women's height in cm), and v005 (the women's sample weight) are other examples of these variables

- **Unused standard recode variables are labeled with "NA" in the variable label**

- **Repeated questions are shown with _ in Stata**
  - Sex of each birth in IR file is denoted as: b4_01, b4_02, b4_03…
  - Place of delivery for multiple births: m14_1, m14_2, m14_3…
  - Variables are in reverse chronological order with _01 (or _1) referring to the most recent birth

- **Missing data can represent:**
  - Variable is not applicable for the respondent
    - In Stata, usually ".", or as "system missing" in SPSS
  - The question should have been answered, but the respondent did not answer it
    - In Stata, usually 9/99/999 vs. ".", but in older data files, all are "."
  - Respondent replied "Don't know" to the question
    - 8/98/998

DHS variables do not contain decimal places. This is because some statistical software programs only keep a few decimal places, and if the variables contained decimal places, results might differ from one software program to another. For example, v005 must be divided by 1000000 before being used as the sample weight.

If you see a variable with a label that contain "NA," it means that question was not asked in this survey.

In the IR file, children are recorded as "repeated questions" variables, with _01 indicating the most recent birth, _02 indicating the second most recent birth, etc. In SPSS, these variables will be in capitalized, with a $ used instead of an underscore, e.g., B4$01, B4$02, B4$03...

OTHER (QUICK) WAYS TO FIND YOUR VARIABLES IN A DHS DATASET

- Scroll through the dataset
- Search through the dataset for keywords in the variables list window
  - In Stata: use the “lookfor” command, e.g.,

    lookfor education

    lookfor “place of delivery”

To find information about women’s education, there are several options—look through the recode manual or scroll through the dataset, both of which are quite large.

A **shortcut** is to use the “lookfor” command in Stata.

Use quotations around phrases you are searching for, but be sure they are phrases in the dataset.

SUMMARY

- Read the DHS Recode Manual to understand the full descriptions of the DHS variables
- For a quick search of the variables you need, use the "lookfor” command in stata
- Chapter 1 of the Guide to DHS Statistics also contains information about the recode file variables.