OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE
www.ophi.org.uk

# Data Issues in Multidimensional Poverty Measurement

Adriana Conconi & Ana Vaz

11 – 23 August 2014
Oxford University, UK

# Outline

1. Sources of multidimensional data
2. Household surveys
3. Indicators' design
4. Applicable population
5. Combined measures
6. Missing values, inconsistencies, "don't know" - Sample drop and bias analysis

# 1. Sources of Multidimensional Data

## Census

- Advantages:
  - information with negligible sampling error;
  - highly disaggregated levels.

- Disadvantages:
  - have low frequency;
  - offer information on a small set of indicators;
  - micro data may not be available to researchers.

# 1. Sources of Multidimensional Data

## Administrative Data

- Advantages:
  - cover virtually all population and in a continuous form;
  - no data collection costs; and
  - data for individuals who might not respond to surveys.
- Disadvantages:
  - information is limited and may not match the research purpose;
  - any changes in data collection procedures or definitions may affect comparability over time;
  - serious data quality issues may compromise accuracy;
  - metadata is usually not available;
  - access to administrative (micro) data varies by country; and
  - linking data sources is rarely straightforward.

# 1. Sources of Multidimensional Data

## Household Surveys

- Most commonly used data source to study poverty
- Collect information on a diverse set of topics on a sample representative of the population of interest
- Areas for improvement:
  - Frequency;
  - Coverage;
  - Dimensional coverage

# Household Surveys: Metadata

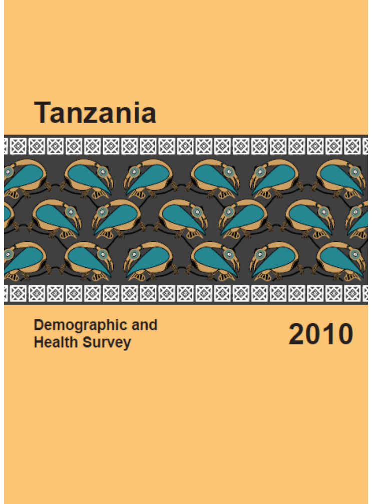
- Metadata is "data about the data".
- Provides information about the survey sample design, fieldwork activities, questionnaires, structure of the dataset, definitions, coding, etc.

**How to use the sample weights?**

**Who are eligible?**

**How to interpret the coding?**

# DHS Country Report

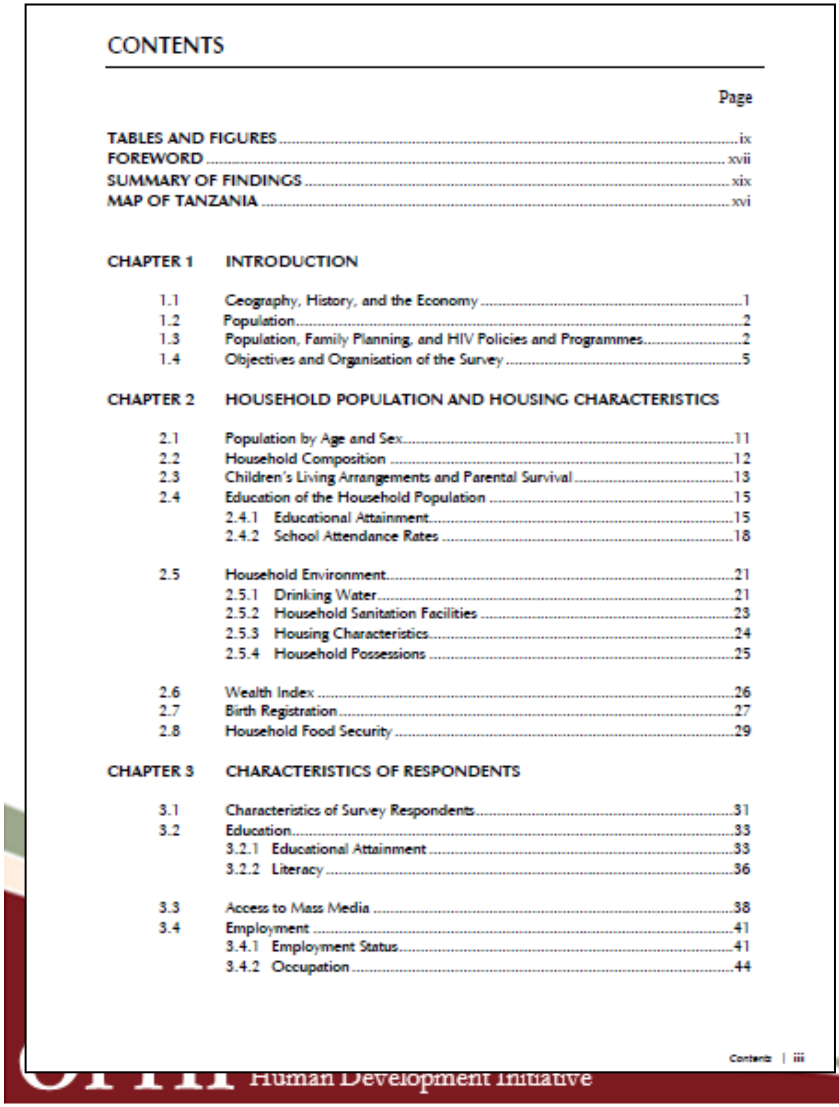


Tanzania

Demographic and Health Survey

2010

**http://www.measuredhs.com/publications/publications-by-type.cfm**



CONTENTS

Page

TABLES AND FIGURES ........ ix
FOREWORD ........ xvii
SUMMARY OF FINDINGS ........ xix
MAP OF TANZANIA ........ xvi

CHAPTER 1 INTRODUCTION

1.1 Geography, History, and the Economy ........ 1
1.2 Population ........ 2
1.3 Population, Family Planning, and HIV Policies and Programmes ........ 2
1.4 Objectives and Organisation of the Survey ........ 5

CHAPTER 2 HOUSEHOLD POPULATION AND HOUSING CHARACTERISTICS

2.1 Population by Age and Sex ........ 11
2.2 Household Composition ........ 12
2.3 Children's Living Arrangements and Parental Survival ........ 13
2.4 Education of the Household Population ........ 15
2.4.1 Educational Attainment ........ 15
2.4.2 School Attendance Rates ........ 18

2.5 Household Environment ........ 21
2.5.1 Drinking Water ........ 21
2.5.2 Household Sanitation Facilities ........ 23
2.5.3 Housing Characteristics ........ 24
2.5.4 Household Possessions ........ 25

2.6 Wealth Index ........ 26
2.7 Birth Registration ........ 27
2.8 Household Food Security ........ 29

CHAPTER 3 CHARACTERISTICS OF RESPONDENTS

3.1 Characteristics of Survey Respondents ........ 31
3.2 Education ........ 33
3.2.1 Educational Attainment ........ 33
3.2.2 Literacy ........ 36

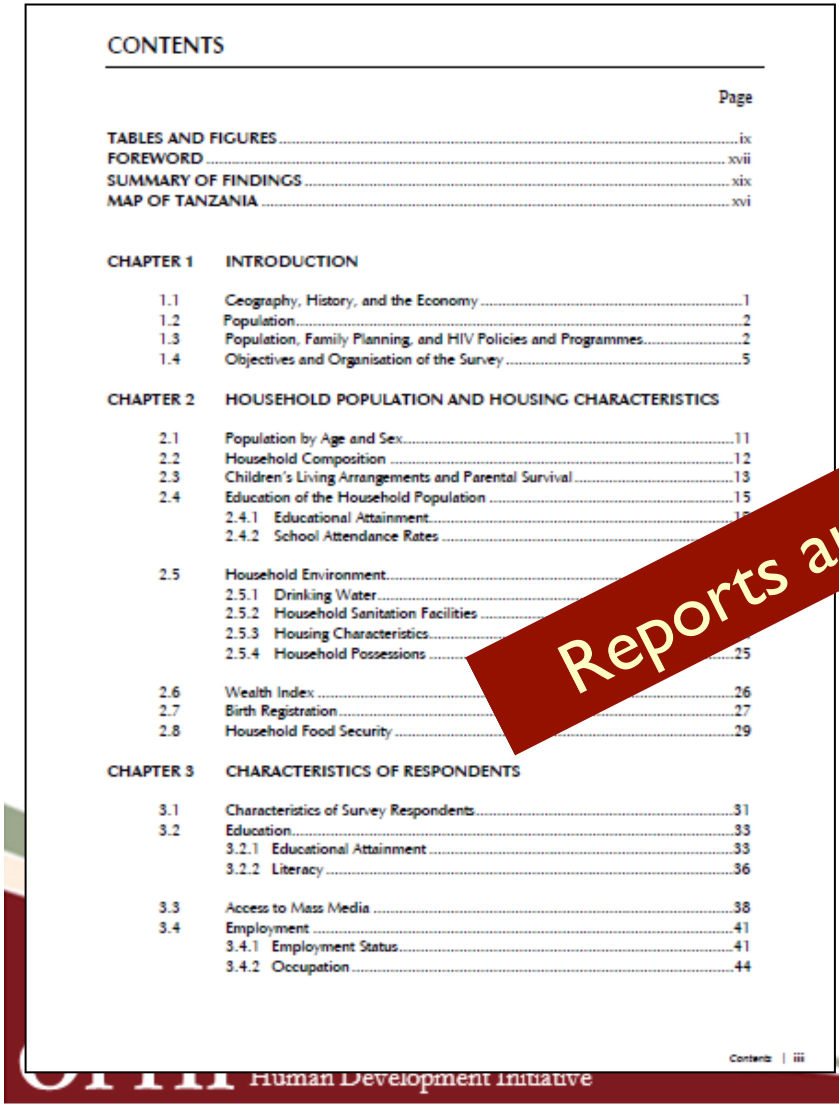
3.3 Access to Mass Media ........ 38
3.4 Employment ........ 41
3.4.1 Employment Status ........ 41
3.4.2 Occupation ........ 44

Sampling design
and representativeness

Tables and results

# DHS Country Report

**http://www.measuredhs.com/publications/publications-by-type.cfm**



CONTENTS

Page

TABLES AND FIGURES ..... ix
FOREWORD ..... xvii
SUMMARY OF FINDINGS ..... xix
MAP OF TANZANIA ..... xvi

CHAPTER 1 INTRODUCTION

1.1 Geography, History, and the Economy ..... 1
1.2 Population ..... 2
1.3 Population, Family Planning, and HIV Policies and Programmes ..... 2
1.4 Objectives and Organisation of the Survey ..... 5

CHAPTER 2 HOUSEHOLD POPULATION AND HOUSING CHARACTERISTICS

2.1 Population by Age and Sex ..... 11
2.2 Household Composition ..... 12
2.3 Children's Living Arrangements and Parental Survival ..... 13
2.4 Education of the Household Population ..... 15
2.4.1 Educational Attainment ..... 1[illegible]
2.4.2 School Attendance Rates ..... [illegible]

2.5 Household Environment ..... [illegible]
2.5.1 Drinking Water ..... [illegible]
2.5.2 Household Sanitation Facilities ..... [illegible]
2.5.3 Housing Characteristics ..... [illegible]
2.5.4 Household Possessions ..... 25

2.6 Wealth Index ..... 26
2.7 Birth Registration ..... 27
2.8 Household Food Security ..... 29

CHAPTER 3 CHARACTERISTICS OF RESPONDENTS

3.1 Characteristics of Survey Respondents ..... 31
3.2 Education ..... 33
3.2.1 Educational Attainment ..... 33
3.2.2 Literacy ..... 36

3.3 Access to Mass Media ..... 38
3.4 Employment ..... 41
3.4.1 Employment Status ..... 41
3.4.2 Occupation ..... 44

Sampli[illegible]
and [illegible]tativeness

Tables and results

Reports are useful but not enough!

# Large Survey projects provide plenty of metadata

**DHS General Data Manuals: available online**

**http://www.measuredhs.com/data/Data-Tools-and-Manuals.cfm**

## Guide to DHS Statistics

Reference to help users who work with DHS survey indicators and datasets to better understand indicator definitions and the calculations used to generate the survey results.

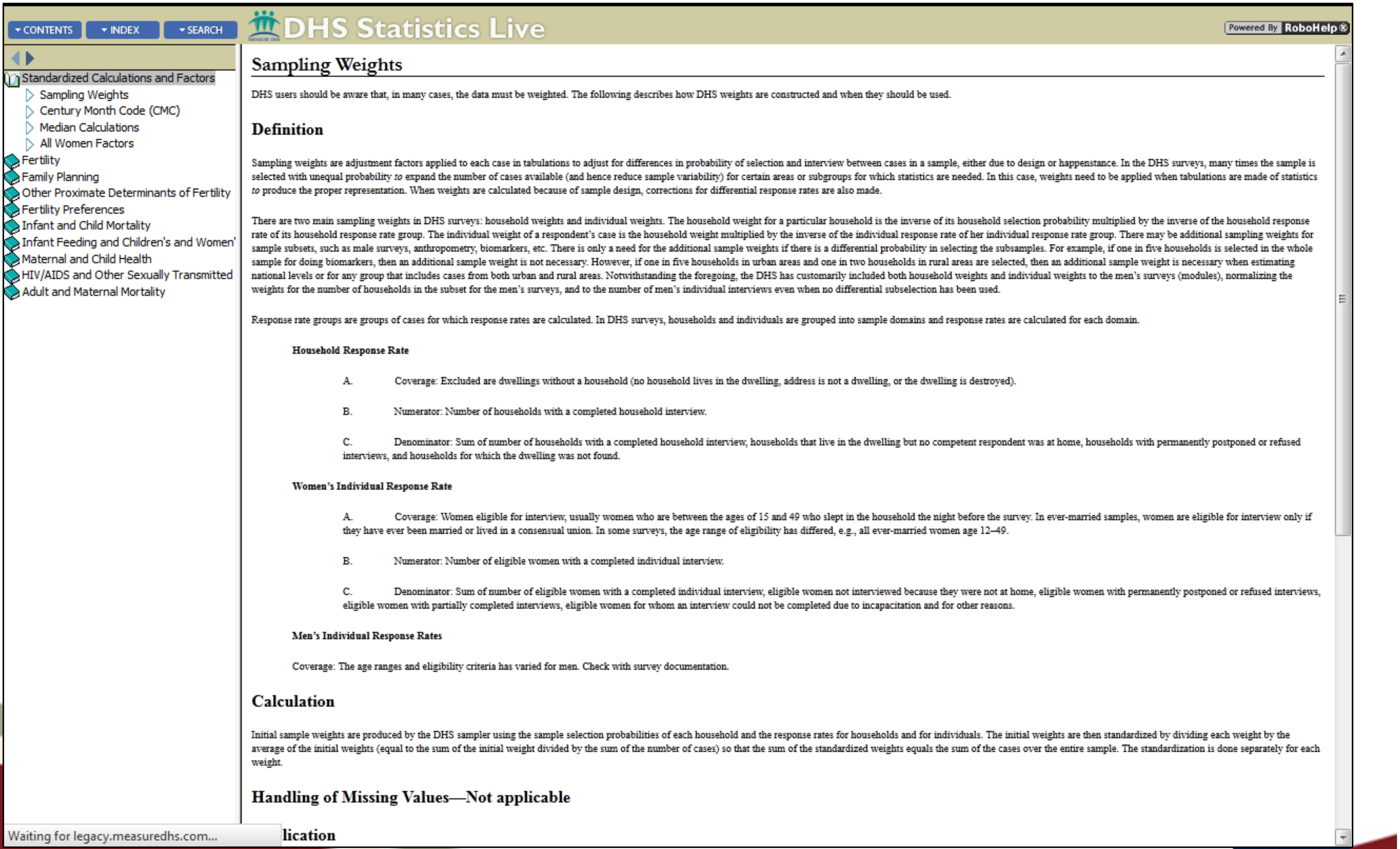
## DHS Recode Manual

Describes each data file and the variables contained in them.

## DHS Data Editing and Imputation

Presents the methodology used by DHS for the production of edited data files. The paper focuses primarily on the editing of dates of events, and the imputation of incomplete dates.

# Online Guide to DHS Statistics – quite handy to search for detailed info!


CONTENTS
INDEX
SEARCH
DHS Statistics Live
Powered By RoboHelp®
Standardized Calculations and Factors
Sampling Weights
Century Month Code (CMC)
Median Calculations
All Women Factors
Fertility
Family Planning
Other Proximate Determinants of Fertility
Fertility Preferences
Infant and Child Mortality
Infant Feeding and Children's and Women'
Maternal and Child Health
HIV/AIDS and Other Sexually Transmitted
Adult and Maternal Mortality
Waiting for legacy.measuredhs.com...

## Sampling Weights

DHS users should be aware that, in many cases, the data must be weighted. The following describes how DHS weights are constructed and when they should be used.

### Definition

Sampling weights are adjustment factors applied to each case in tabulations to adjust for differences in probability of selection and interview between cases in a sample, either due to design or happenstance. In the DHS surveys, many times the sample is selected with unequal probability *to* expand the number of cases available (and hence reduce sample variability) for certain areas or subgroups for which statistics are needed. In this case, weights need to be applied when tabulations are made of statistics *to* produce the proper representation. When weights are calculated because of sample design, corrections for differential response rates are also made.

There are two main sampling weights in DHS surveys: household weights and individual weights. The household weight for a particular household is the inverse of its household selection probability multiplied by the inverse of the household response rate of its household response rate group. The individual weight of a respondent's case is the household weight multiplied by the inverse of the individual response rate of her individual response rate group. There may be additional sampling weights for sample subsets, such as male surveys, anthropometry, biomarkers, etc. There is only a need for the additional sample weights if there is a differential probability in selecting the subsamples. For example, if one in five households is selected in the whole sample for doing biomarkers, then an additional sample weight is not necessary. However, if one in five households in urban areas and one in two households in rural areas are selected, then an additional sample weight is necessary when estimating national levels or for any group that includes cases from both urban and rural areas. Notwithstanding the foregoing, the DHS has customarily included both household weights and individual weights to the men's surveys (modules), normalizing the weights for the number of households in the subset for the men's surveys, and to the number of men's individual interviews even when no differential subselection has been used.

Response rate groups are groups of cases for which response rates are calculated. In DHS surveys, households and individuals are grouped into sample domains and response rates are calculated for each domain.

**Household Response Rate**

A. Coverage: Excluded are dwellings without a household (no household lives in the dwelling, address is not a dwelling, or the dwelling is destroyed).

B. Numerator: Number of households with a completed household interview.

C. Denominator: Sum of number of households with a completed household interview, households that live in the dwelling but no competent respondent was at home, households with permanently postponed or refused interviews, and households for which the dwelling was not found.

**Women's Individual Response Rate**

A. Coverage: Women eligible for interview, usually women who are between the ages of 15 and 49 who slept in the household the night before the survey. In ever-married samples, women are eligible for interview only if they have ever been married or lived in a consensual union. In some surveys, the age range of eligibility has differed, e.g., all ever-married women age 12–49.

B. Numerator: Number of eligible women with a completed individual interview.

C. Denominator: Sum of number of eligible women with a completed individual interview, eligible women not interviewed because they were not at home, eligible women with permanently postponed or refused interviews, eligible women with partially completed interviews, eligible women for whom an interview could not be completed due to incapacitation and for other reasons.

**Men's Individual Response Rates**

Coverage: The age ranges and eligibility criteria has varied for men. Check with survey documentation.

### Calculation

Initial sample weights are produced by the DHS sampler using the sample selection probabilities of each household and the response rates for households and for individuals. The initial weights are then standardized by dividing each weight by the average of the initial weights (equal to the sum of the initial weight divided by the sum of the number of cases) so that the sum of the standardized weights equals the sum of the cases over the entire sample. The standardization is done separately for each weight.

### Handling of Missing Values—Not applicable

lication

# MICS Metadata: available online

**http://www.childinfo.org**

## Country reports

It includes comprehensive survey results and country survey specificities.

## Questionnaires

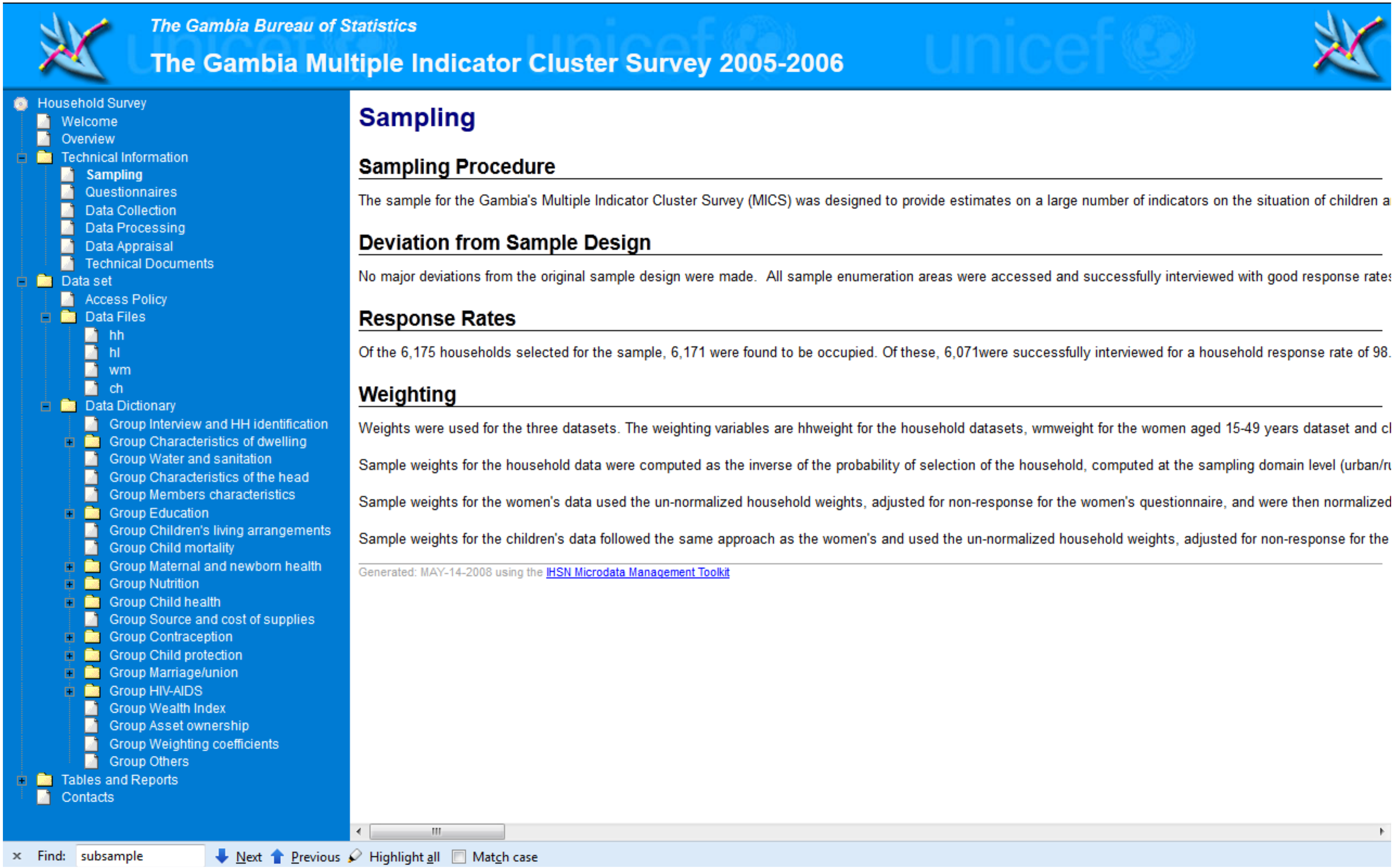
Flow of questionnaire modules, Household questionnaire, Women's questionnaire, Children under-5 questionnaire, Additional situation specific modules, Optional modules

## MICS Manual

## Other various documents

ISample Size (Households) Calculation Template, Pictorials for Water and Sanitation Facilities, One-page pictorial on cooking methods using solid fuels. Sample weight calculation.

# Online MICS Survey archive: provides detailed information describing the content, methodology and implementation of the survey



## Sampling

### Sampling Procedure

The sample for the Gambia's Multiple Indicator Cluster Survey (MICS) was designed to provide estimates on a large number of indicators on the situation of children a

### Deviation from Sample Design

No major deviations from the original sample design were made. All sample enumeration areas were accessed and successfully interviewed with good response rates

### Response Rates

Of the 6,175 households selected for the sample, 6,171 were found to be occupied. Of these, 6,071were successfully interviewed for a household response rate of 98.

### Weighting

Weights were used for the three datasets. The weighting variables are hhweight for the household datasets, wmweight for the women aged 15-49 years dataset and cl

Sample weights for the household data were computed as the inverse of the probability of selection of the household, computed at the sampling domain level (urban/ru

Sample weights for the women's data used the un-normalized household weights, adjusted for non-response for the women's questionnaire, and were then normalized

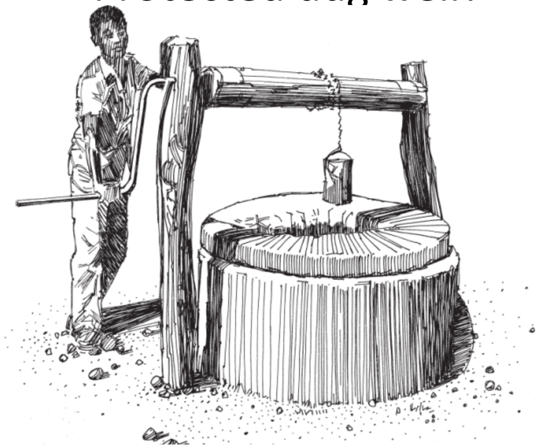
Sample weights for the children's data followed the same approach as the women's and used the un-normalized household weights, adjusted for non-response for the

Generated: MAY-14-2008 using the IHSN Microdata Management Toolkit

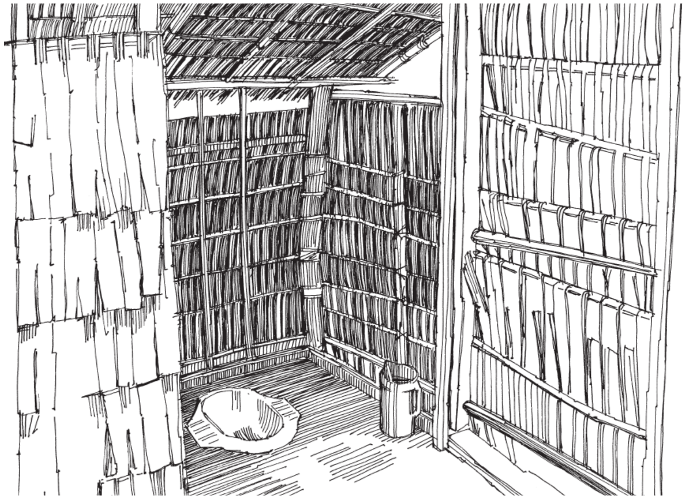
# Search also among materials used during fieldwork

**Example:** Pictorial illustrations on access to water supply and sanitation facilities for use in national household surveys by JPM and UNICEF

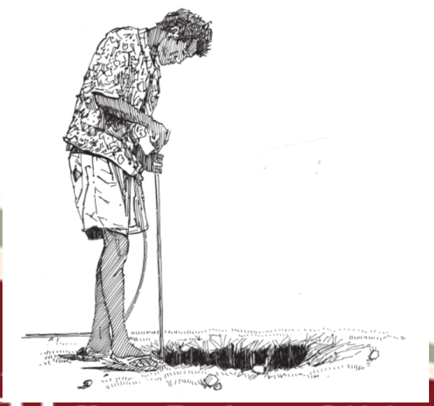
Protected dug well?



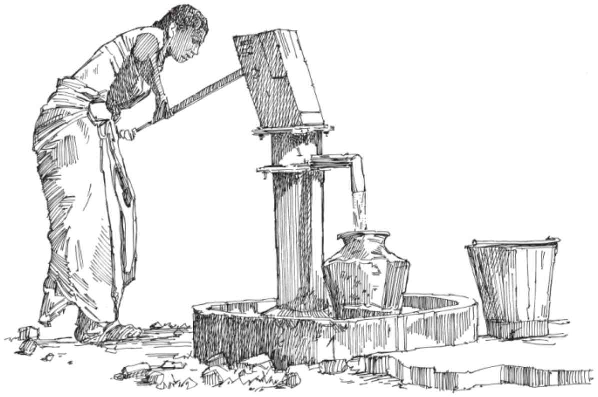
Pour-flush to pit?



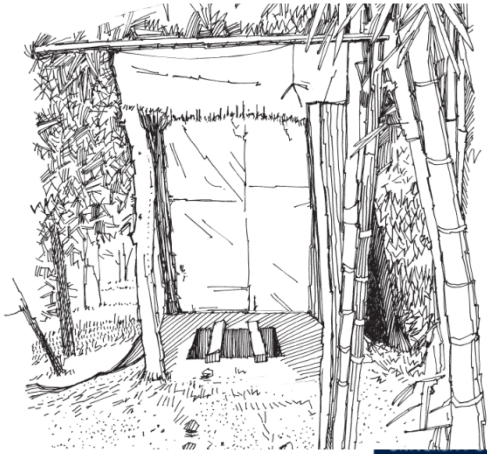
Unprotected dug well?



Tubewell/borehole?



Pit latrine without slab?



# 2. Household Surveys: Survey Design

Usually household surveys follow a complex sampling design in two stages:

1. Clusters (e.g. PSU) are selected from within each strata (e.g. Region+urban/rural)

2. Households are selected from household listings within each cluster (listing are generated during census operations)

The result is a representative and yet efficient sample that reduces cost and increases quality of data collection

# Sample weights

Weights are computed as the inverse probability of selection:

- probability of selecting the cluster;
- probability of selecting the household within the cluster;
- they may also adjust by response rate and/or by the demographic structure of the population (Yansanhe, 2005).

Ignoring the weights would produce significantly biased results

# Samples and subsamples

Some data can be particularly more difficult and expensive to collect, either because it takes longer (e.g. revisits) or it requires enumerators with more expertise (hence supervision is more difficult).

**Check the metadata for subsamples and how to undertake analysis with it!**

*"Changes over Time (regarding Children and Women's Nutritional Status)*

*In phases of the DHS survey before phase IV (DHS+), only children of interviewed women and who were under five years old (or the cutoff for the health section of the individual questionnaire) were weighed and measured. In many surveys, only a subsample of these children were selected for anthropometry. All comparisons between surveys, either over time or between countries, should take into account the possible differences in the defined population base"*

# What geographical level can you decompose?
# Are all ethnic group represented well in the sample?

Survey representativeness depends on the sample design, and will limit how far one can undertake decomposition analysis.

For example in Tanzania:

*"The 2010 TDHS sample was designed to provide estimates for the entire country, for urban and rural areas in the Mainland, and for Zanzibar.*

*For specific indicators such as contraceptive use the sample design allowed the estimation of indicators for each of the then 26 regions. To estimate geographic differentials for certain demographic indicators, the regions of mainland Tanzania were collapsed into seven geographic zones. Although these are not official administrative zones, this classification is used by the Reproductive and Child Health Section of the MoHSW. Zones were used in each geographic area in order to have a relatively large number of cases and a reduced sampling error."*

# Non-response rate and other non-sampling error

It is also a good practice to report the non-response rate and any other non-sampling error (e.g. problems during fieldwork logistics)

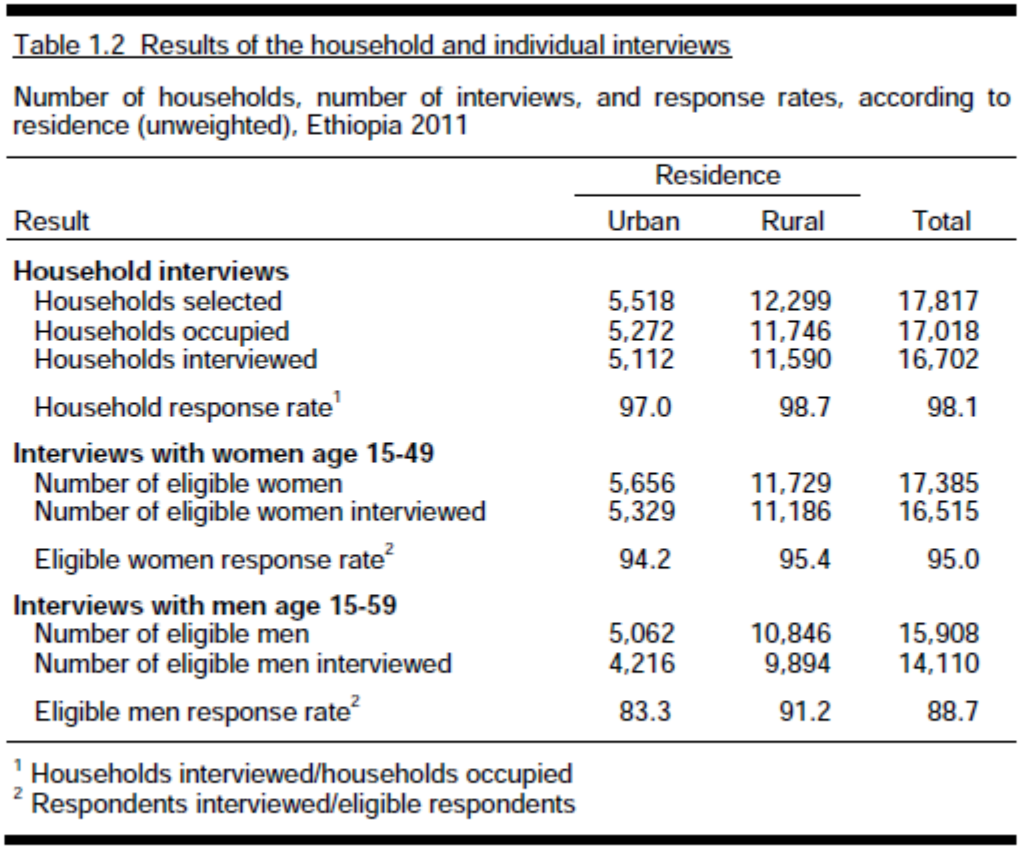
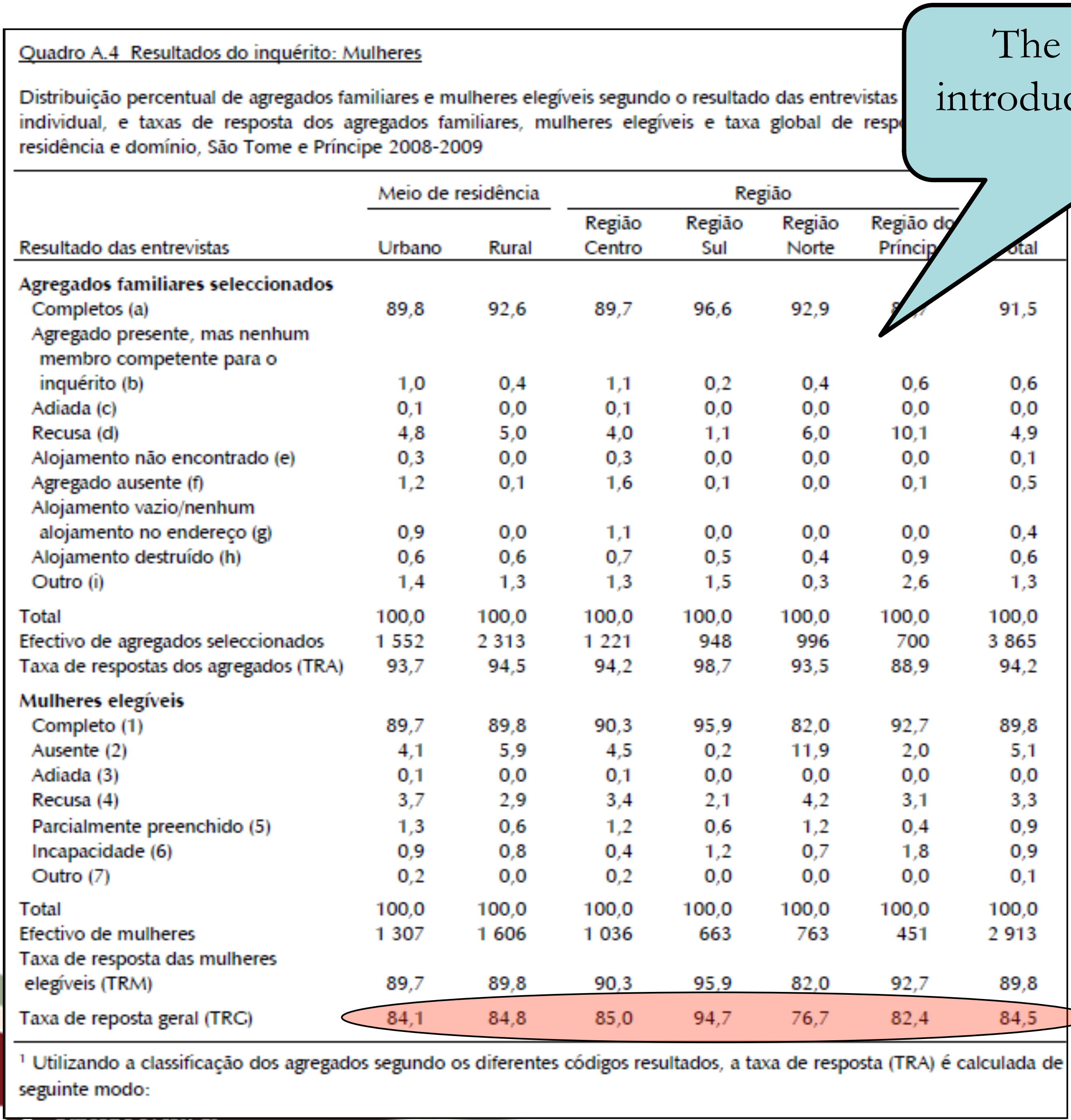


Table 1.2 Results of the household and individual interviews

Number of households, number of interviews, and response rates, according to residence (unweighted), Ethiopia 2011

| | Residence | | |
|---|---|---|---|
| Result | Urban | Rural | Total |
| **Household interviews** | | | |
| Households selected | 5,518 | 12,299 | 17,817 |
| Households occupied | 5,272 | 11,746 | 17,018 |
| Households interviewed | 5,112 | 11,590 | 16,702 |
| Household response rate[1] | 97.0 | 98.7 | 98.1 |
| **Interviews with women age 15-49** | | | |
| Number of eligible women | 5,656 | 11,729 | 17,385 |
| Number of eligible women interviewed | 5,329 | 11,186 | 16,515 |
| Eligible women response rate[2] | 94.2 | 95.4 | 95.0 |
| **Interviews with men age 15-59** | | | |
| Number of eligible men | 5,062 | 10,846 | 15,908 |
| Number of eligible men interviewed | 4,216 | 9,894 | 14,110 |
| Eligible men response rate[2] | 83.3 | 91.2 | 88.7 |

[1] Households interviewed/households occupied
[2] Respondents interviewed/eligible respondents



Quadro A.4 Resultados do inquérito: Mulheres

Distribuição percentual de agregados familiares e mulheres elegíveis segundo o resultado das entrevistas individual, e taxas de resposta dos agregados familiares, mulheres elegíveis e taxa global de resp residência e domínio, São Tome e Príncipe 2008-2009

| | Meio de residência | | Região | | | | |
|---|---|---|---|---|---|---|---|
| Resultado das entrevistas | Urbano | Rural | Região Centro | Região Sul | Região Norte | Região do Princip | otal |
| **Agregados familiares seleccionados** | | | | | | | |
| Completos (a) | 89,8 | 92,6 | 89,7 | 96,6 | 92,9 | [illegible] | 91,5 |
| Agregado presente, mas nenhum membro competente para o inquérito (b) | 1,0 | 0,4 | 1,1 | 0,2 | 0,4 | 0,6 | 0,6 |
| Adiada (c) | 0,1 | 0,0 | 0,1 | 0,0 | 0,0 | 0,0 | 0,0 |
| Recusa (d) | 4,8 | 5,0 | 4,0 | 1,1 | 6,0 | 10,1 | 4,9 |
| Alojamento não encontrado (e) | 0,3 | 0,0 | 0,3 | 0,0 | 0,0 | 0,0 | 0,1 |
| Agregado ausente (f) | 1,2 | 0,1 | 1,6 | 0,1 | 0,0 | 0,1 | 0,5 |
| Alojamento vazio/nenhum alojamento no endereço (g) | 0,9 | 0,0 | 1,1 | 0,0 | 0,0 | 0,0 | 0,4 |
| Alojamento destruído (h) | 0,6 | 0,6 | 0,7 | 0,5 | 0,4 | 0,9 | 0,6 |
| Outro (i) | 1,4 | 1,3 | 1,3 | 1,5 | 0,3 | 2,6 | 1,3 |
| Total | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |
| Efectivo de agregados seleccionados | 1 552 | 2 313 | 1 221 | 948 | 996 | 700 | 3 865 |
| Taxa de respostas dos agregados (TRA) | 93,7 | 94,5 | 94,2 | 98,7 | 93,5 | 88,9 | 94,2 |
| **Mulheres elegíveis** | | | | | | | |
| Completo (1) | 89,7 | 89,8 | 90,3 | 95,9 | 82,0 | 92,7 | 89,8 |
| Ausente (2) | 4,1 | 5,9 | 4,5 | 0,2 | 11,9 | 2,0 | 5,1 |
| Adiada (3) | 0,1 | 0,0 | 0,1 | 0,0 | 0,0 | 0,0 | 0,0 |
| Recusa (4) | 3,7 | 2,9 | 3,4 | 2,1 | 4,2 | 3,1 | 3,3 |
| Parcialmente preenchido (5) | 1,3 | 0,6 | 1,2 | 0,6 | 1,2 | 0,4 | 0,9 |
| Incapacidade (6) | 0,9 | 0,8 | 0,4 | 1,2 | 0,7 | 1,8 | 0,9 |
| Outro (7) | 0,2 | 0,0 | 0,2 | 0,0 | 0,0 | 0,0 | 0,1 |
| Total | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |
| Efectivo de mulheres | 1 307 | 1 606 | 1 036 | 663 | 763 | 451 | 2 913 |
| Taxa de resposta das mulheres elegíveis (TRM) | 89,7 | 89,8 | 90,3 | 95,9 | 82,0 | 92,7 | 89,8 |
| Taxa de reposta geral (TRG) | 84,1 | 84,8 | 85,0 | 94,7 | 76,7 | 82,4 | 84,5 |

[1] Utilizando a classificação dos agregados segundo os diferentes códigos resultados, a taxa de resposta (TRA) é calculada de seguinte modo:

The non-response introduces non-sampling error

# 3. Indicators' Design – Unit Level Indicator Accuracy

- **Unit of identification:** entity who is identified as poor or non-poor – usually the individual or the household.

- HH surveys are usually designed to create indicators that are representative of achievements and/or distributions of some population subgroups.

# Unit Level Indicator Accuracy

- Indicators collected with short reference periods and are judged to be accurate 'on average'. Examples:
  - consumption in the last seven days,
  - illness in the last two weeks, and
  - time use in the past 24 hours.

- However achievements may not be accurate at the individual level. And what if…
  - last seven days' consumption included a family wedding,
  - the respondent had a rare and brief bout of the flu,
  - the last 24 hours was a major public holiday.

# Unit Level Indicator Accuracy

- Indicators used for targeting are always required to be accurate at the individual level.

- Multidimensional measures require the joint distribution of deprivations to be accurate on average.
    - Selected indicators ideally balance indicator precision and unit-level accuracy.

- When tracking changes over time in poverty, indicators should reflect individual achievement levels across the relevant period. No distortions due to seasonal effects, or short-term shocks.

# Indicators Transformation to Match Unit of Identification

- Relevant data may be available for individuals, for the household, and for the community
- So, we may need to transform indicators such that they reflect deprivations of the chosen unit of identification.
- Suppose a child poverty measure with children, household and village level data.
  - How do we construct the $n \times d$ achievement matrix?
  - What is the implicit assumption?

# 4. Applicable Population

- Applicable population: group of people for which the achievement is relevant; namely,
  - it can be measured – it is conceptually applicable – **and**
  - it has been effectively measured – data is available.

- Some achievements relevant for poverty measurement are either conceptually or empirically applicable only for certain population groups.

# 4. Applicable Population

- The achievement may be only conceptually relevant for certain groups:
  - Income
  - Vaccinations
  - Employment status

- The achievements may be conceptually applicable to the whole population but data is only collected for some groups…
  - Anthropometric indicators

How to deal with this?

# 4. Applicable Population

- To restrict consideration to universally applicable achievements
  - Narrows the set of indicators

- To construct group-specific poverty measures
  - Discriminating by groups may not eliminate applicability issues
  - Not possible to track national poverty or target households
  - May miss the overlaps of disadvantaged groups

- To combine achievements that are not universally applicable (e.g. Global MPI)

# 5. Combined Measures

- Approach followed when constructing the MPI.

- All household members are deprived:
  - If has at least one child or women undernourished
  - If at least one child in the household died
  - If at least a child of school age is not attending school

- Assumption of negative externalities

- Indicator assuming a positive externality:
  - All household members are considered non-deprived if at least one person has five years of schooling

# 5. Combined Measures

- How to deal with households where not even one person qualifies for the achievement under consideration?
  - Drop these households from the sample.
    - ➢ That would bias the estimates…
  - Drop the indicator and re-weight the remaining indicators
    - ➢ That would violate dimensional breakdown…
  - Consider them as non-deprived (deprived) in that indicator
    - ➢ Need to scrutinize this assumption…

# 5. Combined Measures

- Suppose survey has not collected information from all applicable members. How to deal with households where there is no data for any member?
  - Consider them as non-deprived (deprived) in that indicator

  - Considering them as non-deprived could be seen as a ‘conservative’ approach, and will lead to a `lower bound‘ poverty estimate.

# Assessing Combined Measures

- Potential household composition effect

- Inclusion of indicators referring specific groups can be made provided that:
  - Not all indicators refer to a particular specific group
  - An important proportion of households have at least one member for whom the achievement is relevant
  - Empirical test of the impact of the household composition

# Assessing Combined Measures

- Alkire and Santos (2014)
  - Tests of differences in means between MPI-poor and non-poor households in terms of size, number of children under 5, number of females, number of members 50 years or older, proportion of female-headed households, and proportion of school-aged children.
  - Decompose country's MPI by age and gender and compare the rankings, correlations and proportion of robust pairwise comparisons.

# 6. Missing values, inconsistencies, “don’t know”

**Missing value:** a variable that should have a response, but because of interview errors the question was not asked.

**Inconsistent:** This code is generally used by people in the secondary editing group, when a value or code is not plausible.

**“Don’t know” responses:** These codes are normally pre-coded in the questionnaires, but they are consistently used throughout the recode file.

## How should we treat missing values?

**What to do if we only have a few missing cases when constructing a aggregate household indicator? (e.g. Years of schooling or school attendance)**

# 6. Missing values, inconsistencies, "don't know"

Ways to deal with missing values:

1. Use rule to assign value for the missing data. E.g. Global MPI:
   - Household non-deprived if at least one member has 5+ years of education.
   - Household deprived if we have information for at least 2/3 of the household members and none of them has at least 5 years of education.

2. Drop the observation from the sample. E.g. Global MPI:
   - Household with missing information in any of the relevant indicators are dropped from the sample

# How should we treat a missing value when computing the Adjusted Headcount Ratio?

Suppose the following matrix, and a poverty line of k>=1:

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ . & 0 & 1 \\ 0 & 0 & 0 \\ . & . & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

How do we compute the average deprivation with missing information?

Is this individual poor?

In practice we reduce the sample to only cases with information in all  indicators, having a consequent "sample drop" due to missing information

# Sample Drop and Bias Analysis

**Problem**: sample drop may lead to biased estimates

**Bias analysis**: group with missing values is compared to rest, using the indicators for which values are present for both groups

- Series of hypothesis test for difference of means or proportions

**Results**:

- No significant differences: we can use the reduced sample
- Significant differences: we can still use the reduced sample but should explicitly the direction of the bias

Tabita, Kenya
Rabiya, India
Stéphanie, Madagascar
Agathe, Madagascar
Dalma, Kenya
Ann-Sophia, Kenya
Valérie, Madagascar

# Thank you!

# Sample Drop and Bias Analysis

- **What about using imputation?**
  - Estimate a model with the achievement as the dependent variable against a set of explanatory variables
  - Use estimated parameters t predict achievements  for cases with missing values

- **Limitations**:
  - The estimated model needs to be accurate
  - We would have to specify a model that could predict a vector of deprivations
  - Cannot solve problem of non-applicable populations

**Need further research!**

# Sample drop and bias analysis

The sample drop may…

Affect the representativeness of the sample

- Need to check the proportion of missing values for each indicator and analyze the proportion of total sample drop

Affect the population share when regions are decomposed

- Need to check how the share of each region changes before and after sample drop – Is there a bias towards a particular region?

# Sample drop and bias analysis

In practice this "bias analysis" can be undertaken with a series of hypothesis test for difference of means or proportions.

Often the cause of a large sample loss is only few indicators. So one could assess if the subsample after sample drop has a significant bias in any given indicator.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Where $\mu_1$ represents the estimation for the full sample, and $\mu_2$ represents the estimation for the subsample after sample drop