

# LIS Policy on the Treatment of Missing Information

August 2007

## **TABLE OF CONTENTS**

I. Introduction .....	1
II. LIS Treatment of Missing Information .....	2
A. Waves I through IV .....	2
1. Monetary variables .....	2
2. Non-monetary variables .....	2
B. Wave V .....	2
1. The universe of LIS variables .....	3
2. Missing values in LIS variables .....	3
3. Not in the universe.....	3
a) Type 1: Monetary variables.....	3
b) Type 2: Non-monetary variables .....	4
c) Examples .....	4
III. Special Issues .....	4
A. Panel Surveys.....	4
1. Observations included in the LIS cross-sectional data file.....	5
2. Shadow files .....	5
3. Dropped observations .....	5
B. Child Files .....	5
1. Waves I through V.1 .....	5
2. Wave V.2.....	6

## **I. Introduction**

LIS uses both survey and registry data to create the LIS variables. Information from both types of data bases may be missing for a variety of reasons.

Missing information in surveys can occur, for example, if a respondent is unwilling to answer the survey or certain questions in the survey, or if he does not know the answer to a survey question. In others cases, some household members may be unavailable to answer questions due to travel or ill health. Additionally, the survey method could also lead to missing answers. For example, if the interview was carried out using a paper questionnaire, some answers may be illegible or a block of questions may be skipped inadvertently or through misrouting.

In registry data, information could be missing if no taxes were owed (e.g., because of no employment or unpaid work) or if the work and the payment were in different reference period (e.g., wages are paid in January for work performed in December of the previous year).

The amount of missing information varies across countries and years, and may be substantial.

There are three main types of missing information of concern to LIS:

1. The question was asked of the respondent but he or she was unwilling or unable to answer.
2. The question is not applicable to the individual.
3. The question is asked only of a specific subgroup, and as a result certain respondents were not offered the possibility of answering (i.e., there is a skip pattern in questionnaire or an individual is not included in the register).

Each survey provider handles missing information in its own way. In some surveys, the data collection institute chooses to eliminate missing information through imputation. In these cases, missing data are overwritten through a process that may be based on available answers from “similar” respondents. Some providers code the missing answers with information about the reason the information is not available. Some choose to leave the answers as missing.

Faced with a multitude of reasons for and responses to missing values among countries, LIS has established a policy to help consistently deal with missing information. This document summarizes the general methods for dealing with missing values prior to Wave V and lays out the framework for dealing with missing values in the LIS variables as of Wave V.2.

## **II. LIS Treatment of Missing Information**

### **A. Waves I through IV**

#### **1. Monetary variables**

The coding of missing information in LIS data was not standardized in data sets up through Wave IV. The only fixed rule that was followed was to code monetary variables (e.g., income, expenditure) as zero (0) if no amount was available. As a result, the meaning of the 0 was not uniform. In some cases, the zero value means zero value; in others stands for “amount unknown”.

#### **2. Non-monetary variables**

Following the coding within individual surveys, the non-monetary variables in earlier waves were often given separate codes to express reasons for missing information (e.g., “unknown”, “refusal”, “question not asked”, “not in register”). While this practice has advantages, the codes themselves were unfortunately not standardized, but varied across variables and countries.

### **B. Wave V**

Starting from LIS Wave V.1, a standardized way of coding missing information was adopted for all variables. The rules applied depend on both the reason the information is missing (e.g., refusal or not applicable) and the type of variable (i.e., monetary or non-monetary).

## 1. The universe of LIS variables

The rules adopted by LIS to classify missing information are based on whether the unit (household or person) should be “in the universe” of the LIS variable. A unit belongs in the universe if the information in the LIS variable is applicable to the unit.

In most cases, identifying the universe for a LIS variable is fairly straightforward. For example, if a LIS variable is based on one original survey question that is intended for all adults, then all adults are in the universe of the question whether they refused to answer or were never asked due to absence from the household or other reasons.

When a LIS variable is constructed from more than one original survey question, however, the universe may be more difficult to determine. This is especially true when each of the original survey questions is asked of different individuals (e.g., only the employed and unemployed are asked for the same information, but through different questions). This situation primarily arises in the construction of the labour market variables beginning in Wave V.2 and is discussed in further detail in the “Guidelines for Labour Market Variables” document (<http://www.lisproject.org/techdoc/labourmarket.pdf>).

The universe and reference period for all variables in all countries is provided in the country-specific documentation.

## 2. Missing values in LIS variables

A LIS variable is coded as “system missing” (i.e., “.”) if the unit of observation is in the universe of the LIS variable (as described above), but the individual (1) refused to answer the question; (2) responded that they did not know or could not answer the question; or (3) was not asked the question for other reasons.

## 3. Not in the universe

The units of observation that are not in the universe of the LIS variable are coded in one of two ways, depending on the variable type. LIS divides variables into two types: (1) monetary variables; and (2) non-monetary variables.

### a) Type 1: Monetary variables

Monetary variables include income, expenditure, social benefits, and LIS income aggregates.

In order to maintain consistency with earlier waves, monetary variables are coded as zero (**0**) if the question is not asked or does not pertain to the unit of observation. In most cases, a zero value will reflect a zero amount (e.g., if the question about the amount of child benefits “does not pertain”, then the individual received 0 benefits).

While clearly defining the universe decreases the possibility of a zero value standing for “amount unknown”, if a major block of individuals is not asked a

question, the zero value may be standing in for a missing value (e.g., a child with positive income). Pay attention to the universe of the LIS variable (and related variables)! By restricting your sample to a subset of the universe, you can avoid inadvertently including zero values that would bias your results.

**b) Type 2: Non-monetary variables**

Non-monetary variables include all other LIS variables (e.g., demographics and labour market). In designating observations as “not in the universe”, LIS treats all non-monetary variables in the same manner. Missing observations are coded as **-1** if the individual belongs to a major group that was not asked the original survey question (e.g., children or the retired), or the question does not pertain to the unit of observation (either as stated by the individual or imputed by the survey provider).

Further, in the case of the labour market variables regarding hours and weeks worked, LIS imputes 0 values when it is known that an individual was not employed over the reference period. In some cases, these values can not be imputed for everyone. Please check the country-specific documentation carefully to determine the universe of each variable.

**c) Examples**

In order to illustrate the difference between Type 1 and Type 2 variables, here are a few examples:

**(1) A person who is not employed**

Occupation and industry (Type 2) will be assigned the value of -1. Salary received (Type 1) will be set to 0.

**(2) A single-person household**

In a single-person household, there is no spouse present. Therefore, the age of the spouse (Type 2) is -1, whereas the amount of the pension of the spouse (Type 1) is coded 0.

**III. Special Issues**

**A. Panel Surveys**

Special issues arise when using panel data in cross-sectional analysis. This section explains how LIS deals with the available information in the relevant cross-section of panel data sets. Note that while the source data for some countries does, indeed, stem from panel surveys, the LIS identifiers can not be used to link the same persons or households across waves. All LIS data are treated as cross-sectional data even if they originate from a panel survey.

## **1. Observations included in the LIS cross-sectional data file**

LIS defines three distinct samples within the household-level panel data set: the **cross-section**, the **wave**, and the **roster**. The cross-section can be described as those household that belong to the representative cross-section (i.e., those households with positive cross-sectional weights). The wave includes all those households and/or individuals who were interviewed during the wave of the cross-section, regardless of household weight. Finally, the roster includes any individual or household that was ever interviewed during the course of the panel.

The primary LIS data sets (household and person) include the cross-section, which is defined as all households and all individuals with positive household weights. Individuals who were not interviewed are included with missing values for all unanswered questions. In many cases, LIS-aggregated variables for households with missing person-level information are also missing (e.g., DPI).

## **2. Shadow files**

In some surveys, individuals and/or households were interviewed even if they were not included in the representative cross-section. (These include, for example, an oversampling of new immigrants starting in a late wave in the German panel, and split-off households in Russia.) LIS refers to these as the wave observations that are not included in the cross-section. These observations have valid values for the survey and LIS variables, but have zero or missing cross-sectional household weights. Rather than discard this information, LIS has included these observations in a “shadow” file for the relevant country and year. (See country-specific documentation for shadow files for details of their use.) While they are not part of the representative cross-section, the information could be of interest to researchers concentrating on a specific sub-sample (e.g., recent immigrants or young families) where weights are not required in the analysis.

## **3. Dropped observations**

Individuals in the roster who do not have valid information for the relevant LIS variables and who have zero or missing weights are dropped and are not included in any LIS data set.

## ***B. Child Files***

### **1. Waves I through V.1**

Prior to Wave V.2, LIS provided two separate files at the person-level: a person file (individuals 15 years old and above) and a child file (14 and under). If information on children was included by the data provider, the full set of LIS variables was calculated in the person-level file for all individuals 15 years of age and older. For children (14 and under), only cursory information was provided (person identifiers, relationship to the household head, age, age rank among siblings, gender, and cross-sectional weight).

As of the release of Wave V.2 in February 2007, ***the child file ceases to exist for all data sets in all waves.*** Information in the child file has been appended to the person file for Waves I through V.1. Where the child file existed in previous waves, children appended to the person file are assigned a value of **-2** for all variables included in the person file that were not in the child file for these waves (i.e., all variables except country, casenum, prel, page, psex, ppnum, and pweight).

These variables were assigned a special “missing” value (-2) in order to distinguish them from the true missing values and those “not in the universe” as of Wave V.2. In some cases, information for those 14 and under was given by the data provider, but was not included because of the child file age cut-off. In other cases, children were included in the original data, but the data provider did not ask questions to children (defined at different ages, depending on the country). In still others, the data provider only gave information on adults.

Rather than investigate the “universe” for each and every variable in previous waves, LIS has chosen to assign the special value of -2 to notify users that these values may be a true missing or it may be that those individuals are not in the universe. In so doing, LIS is providing the same file format for all country-years (i.e., one file for all person-level information) in order to ease the programming requirements of LIS users.

Please note that if data sets from earlier waves are modified for other reasons, LIS will follow the current rules for missing values, and the -2 value will no longer appear in the re-lissified data.

## **2. Wave V.2**

As of Wave V.2, children are treated in the same manner as adults when information on children is reported by the data provider. If children are “in the universe” of the LIS variable (as defined above), they are given a valid or missing value. If they are “not in the universe”, they are assigned a value of 0 or -1, depending on the variable type.