

WageIndicator



in Argentina, Belgium, Brazil, Denmark, Finland, Germany, Hungary, Italy, India, Mexico, Netherlands, Poland, Spain, South Africa, South Korea, United Kingdom, United States

[www.wageindicator.org](http://www.wageindicator.org)

## **THE *WAGEINDICATOR* WEB-SURVEY: DATA STORAGE AND DATA CLEANING**

Kea Tijdens, AIAS/University of Amsterdam, Netherlands

15/01/2007

### **Data storage**

The data is stored on a special, secure website, accessible to the data-manager only. The data is stored as delimited text. The data-manager taps the data usually once a week, and converts the delimited text into a SPSS statistical file. Until November 2005, only fully completed questionnaires were stored, i.e. from the respondents that had pressed the button SEND at the end of the questionnaire. From November onwards, incomplete questionnaires are also stored.

The data-management company is also responsible for the content of the QMS. Whatever failures are detected in the dataset, they can be repaired in the QMS instantly. Vice versa, any changes in the QMS are communicated instantly with the person responsible for the dataset.

### **Anonymity of the respondents**

As for the anonymity of the respondents, the *WageIndicator* web survey does not ask for names or addresses of individual participants, or any other direct identifier. Respondent-side IP numbers are not registered.

The *WageIndicator* web survey asks for email addresses to notify the prizewinner and to invite respondents to complete the survey again one year later, if they have indicated that they want to do so. Privacy of participants is safeguarded, because the email addresses are separated from the remaining data immediately and they are stored on a separate computer. The email addresses will never be made available to neither any national *WageIndicator* team nor to any Third Party. This is promised the participants when they complete the question in which their email address is asked. The dataset does not include email addresses. Panel members are identified by their unique identifier number.

## Quarterly releases

The continuous web survey is released on a quarterly basis. The Table with the start and end dates shows some overlapping days. This is related to the fact that the data comes from three servers.

*Start and end dates of the quarterly values*

release	start date	end date
1	30.09.2004	05.01.2005
2	06.01.2005	14.03.2005
3	14.03.2005	04.07.2005
4	04.07.2005	09.09.2005
5	01.09.2005	30.12.2005
6	30.12.2005	31.03.2006
7	31.03.2006	16.06.2006
8	17.06.2006	30.09.2006

## Data cleaning for multiple responding

Per release the data-manager performs several tests to trace multiple responding. First, it may happen that for technical reasons a respondent is more than one time present in the dataset. This is identified as having the same values for all variables. Multiple cases are removed.

Second, the data is controlled for respondents who have intentionally completed the questionnaire multiple times, for example because they want to increase the likelihood of winning a prize. This is the case for observations with the same value on the sum of the values for region of work, industry, occupation, year of birth, gender, education, presence of children and gross wage. Multiple cases are removed.

## Data cleaning for incomplete questionnaires

Per tap the data-manager performances tests for technical failures. An observation passes this test when having valid values for six critical variables, notably education, occupation, industry, wage, sex, and year of birth. Observations that do not pass this test are deleted for the final dataset.

Per tap the data-manager performances tests as for incomplete questionnaires. The dataset only has the completed questionnaires. The data of the incomplete questionnaires are available on request. From November 2005 until December 2006, the definition of incomplete was anybody who had not ticked the final **SATLIFE** question. Since January 2007, however the incomplete questionnaires were defined as those respondents who had not ticked any item from the question about children **CHLD** onwards. The reason was that in November 2006, due to a technical failure, the data of **SATLIFE** for a thousand respondents were lost, but they were nevertheless included in the quarterly dataset.

## Panel members

Panel members are recruited using a question whether they are willing to complete next years' questionnaire. After a year, they are invited by email to complete the questionnaire again. In the dataset, they can be identified by their identification number **idkey**, which is the same as a year earlier.

## Out of range values

Out of range values seem to be impossible in a web-survey. Nevertheless, they occur, because the textboxes sometimes cause problems. In case the respondents have typed a semicolon or a hard return in the open-ended questions, they are interpreted as a separator when converted from the txt-file to the SPSS-file, causing out-of-range values in the successive variables. In every release, this appeared to be the case in less than 0.1 percent of the cases.

From release 8 onwards, the problem is solved because it became possible to prescribe the order in which the delimited text data was gathered from the server. The data-manager could define the sequence of the data, irrespective of the sequence of the questionnaire. Now, the data could be grouped by variable formats, placing the text boxes at the end of the sequence.

## Missing values

The Table shows that the QMS assigns system missing values to three categories of questions that generate no data. Per release the data-manager assigns user-missing values to some initial system missings.

### *System missings in the stored txt data*

Questions with initial system missings	in the dataset
Questions that are not shown to the respondents, because they are switched off for the country, for the category or because it is a short questionnaire	system missing
Questions that are not shown to the respondents, because of the routing	user missing
Questions that are shown to the respondents, but they have not ticked a response	user missing

The full list of user missing values is shown in the Table. In the QMS, all response categories 'Not applicable' and 'I don't know' are assigned value -8 respectively -7. Out-of-range values were initially recoded system missing, but since release 8, they are coded -3. Since January 2007, the questionnaire has randomized items. If an item in the item pool is not shown to the respondent because of the randomization, the variable has the value -6.

### *Missing values in the dataset*

Value	Label
-9	User missing
-8	Not applicable
-7	I don't know
-6	Not asked (random item)
-3	Out of range
-1	Not asked (skipped)

## Data cleaning

For every release, a series of cross-tabulations of almost all variables by country by release are run, whereby all missing values are set to values. Here, unexpected variations across releases and across countries are checked. In addition, when all missing values are set to missing values again, averages are checked across releases and across values. Unexpected averages are given a closer look, by a cross tabulation of the variable across survey weeks in the release. It may very well be the case that for technical or other reasons there was no data-intake in a week.