GENDER - General Target Variable Report (GVR)

1. General Information

The target variable T_GENDER indicates the gender of a respondent. It takes a value of 0 for males, 1 for females, and 2 for individuals who identify themselves in other ways (see Table 1.1).

The target variable report for T_GENDER is accompanied by the following Excel documents:

- The Detailed Variable Report (DVR): T_GENDER_DVR_SDR2.xlsx.
  DVR Excel files in SDR2 systemize all information about source variables that were used for harmonization into a given target variable of the SDR2 database;
- The CrossWalk Table (CWT): T_GENDER_CWT_SDR2.xlsx.
  CWT Excel files in SDR2 contain details about mapping of source values to target values.

Table 1.1. GENDER: Description of the Target Variable

	Variable description	**Variable name**	**Variable values** *
Target variable	Respondent's gender	T_GENDER	0 = Male 1 = Female 2 = Other
Source variables			See: T_GENDER_DVR_SDR2.xlsx T_GENDER_CWT_SDR2.xlsx
Control variable	Quality control: Processing error types (realized values)	QP_GENDER	0 = No processing errors 3 = Contradictory variable values 5 = Lack of variable value labels

* Missing values are assigned according to the SDR2 missing codes schema, provided in the Appendix.

2. Survey Projects

Source variables that we used for T_GENDER appear in all 23 international survey projects, its 174 waves and all but five national surveys – 3324 out of a total 3329 national surveys. The data cover 156 countries and years from 1966 to 2017. Five national surveys that lack data on gender are: EB/77.3 for Montenegro and Serbia, WVS/1 Finland, Mexico, and South Africa.

3. General Rules and Procedures

3.1. Source data description

The scheme of sex/gender coding in source data, where 1 is male and 2 is female, appears in all 23 international survey projects. Exceptions are LITS/1, where male is coded as 1 and female as 0, and NEB/3 and NEB/7, where male is coded as 0 and female as 1.

The question on gender does not often appear in the questionnaire. When it does, the wording in the English language master documents varies, with some questions asking about gender and others about sex. Generally, there is no sufficient information to understand whether data about gender are provided by the interviewer, coordinator or respondent. In some cases (e.g. CB/2008) it is clearly documented that gender information was derived from the sampling procedure.

In the case of one national survey – AMB/2016 Canada – the source data include a value for gender other than male or female.

For details on source variables pooled to create this target, see its detailed variable report T_GENDER_DVR_SDR2.xlsx

3.2. Rules of transformation of source variables into target variable

We recoded variables so that the code '0' corresponds to male, "1" corresponds to female, and "2" corresponds to individuals identifying themselves in other ways. Although the "other" gender category is present in only one national survey, we decided to include it as a target value category.

Missing values and different situations that warrant to be treated as missing data are coded according to the SDR2 missing codes schema, provided in Table A.1 in the Appendix.

3.3. Methodological variables that accompany T_GENDER

The target variable T_GENDER is accompanied only by a quality control variable for processing error types, QP_GENDER.

QP_GENDER, stored in the SDR2 PLUG_SURVEY file, measures processing error types that we identified when comparing information about the source variables available in the survey documentation (codebook, questionnaire) with information about the same source variables in data records in the source data files (data dictionaries). Put differently, QP_GENDER accounts for inconsistencies between different metadata elements, such as data records on the one hand (i.e. variable values, variable labels, value labels – in the source data files), and information in the codebook or questionnaires, on the other hand.

Generally, in the SDR2 project we check for five types of processing errors that we define below:

- **Illegitimate Variable Values** captures values of the source variable that are outside of the range that SDR proposes as acceptable for a target variable's value range.
- **Misleading Variable Values** shows that variable values, as coded in the questionnaires/codebooks, are not congruent with the data records in the source data files.
- **Contradictory Variable Values** captures inconsistencies in how the same variable value is labeled in a codebook vs. a questionnaire or a data dictionary.
- **Variable Values Discrepancy** concerns situations when several or all values in the same variable are inconsistently labeled in codebook, vs. questionnaire or a data

dictionary (i.e. there are reversed scales or contradictory value labels for more than one variable value).

- **Lack of Variable Value Labels** concerns undefined "nulls" and codes (variable values) in the source data file that are not explained in any source of documentation defining variables and their values (i.e. "wild codes" for a given variable).

Correspondingly, the full list of SDR2 codes for Processing Error Types is:

0 = No processing errors
1 = Illegitimate variable values
2 = Misleading variable values
3 = Contradictory variable values
4 = Variable values discrepancies
5 = Lack of variable value labels
6 = Illegitimate & misleading variable values
7 = Illegitimate & contradictory variable values
8 = Illegitimate & discrepancy variable values
9 = Illegitimate & lack of variable value labels
10 = Contradictory variable values & lack of variable value labels
11 = Discrepant & lack of variable value labels
12 = Illegitimate & contradictory & lack of variable value labels
13 = Illegitimate & discrepant & lack of variable value labels

Values 6 to 13 indicate that a combination of different processing error types occurs within one source variable.

Source variables used to create T_GENDER contain two out of the five processing error types discussed earlier: 3 = contradictory variable values, and 5 = lack of variable value labels. There are no instances when more than one processing error type appears in the same source measure of gender. Thus, QP_GENDER takes the values 0, 3 and 5 from the full list of SDR2 codes for Processing Error Types presented above.

4. Special Cases

Availability of source variables:

- At the level of national survey (project/wave/country), information about respondent's gender is not provided in EB/77.3 for Montenegro and Serbia, WVS/1 Finland, Mexico, and South Africa. For participants of these surveys, T_GENDER takes the SDR2 missing code QNA (see Table A.1 in Appendix)
- AMB/2014 and AMB/2016 provide two variables with identical frequencies: "q1" and "sex". We chose "q1" to be consistent with the previous AMB waves.
- AMB/2012 provides two variables concerning gender: "q1"and "sexo". We chose "q1", as "sexo" had suspicious frequencies ("HOM 574, MUJ 997, N/A 40,061").
- In CNEP/4 for Indonesia there were two source variables concerning respondents' gender with similar labels and descriptions in documentation: L.GENDER and L.GENDER.F. We chose L.GENDER.F as L.GENDER had only missing values.

Documentation:

- The variable genderB from LITS/1 lacks labels for variable values in the in data records in the data files (data dictionaries). We identified this source variable as potentially matching our target on the basis of the variable name itself and technical reports. LITS/1 technical report included labels for values used in genderB and distributions, which made it possible to map LITS/1 genderB source values into our target.
- In many instances, questions concerning gender are missing from questionnaires, and so is the wording of the instruction for the interviewer to determine gender. However, most codebooks contain information about this variable (usually without the question wording, but with basic coding information). It is difficult to establish in which instances gender was determined by an interviewer and in which it was asked directly, even if the question about gender is present in the questionnaire.

Sources of information concerning respondents' gender:

- In CB/2008 the information concerning gender of respondents was based on sampling procedure (cf. the label in the source data file: "Sex from Kish Table [R]"). (Source: CB/2008 SPSS dictionary.)
- In national surveys within ISSP waves 2004-2015 in some cases information about respondents' gender was taken from (or supplemented by) other sources described in ISSP codebooks, i.e. in Norway "Item non-response replaced by data from the Central Register of Persons", in Spain "Question derived from National Register", in South Africa "Copy from contact sheet", in Japan "Respondent's sex can be known also from the sampling frame", in China "Information automatically taken from sample database", in Iceland "Sex is one of the variables that come with the sample list from the National Registry of Iceland. The sample information was used for those who did not give an answer about their sex. Therefore there are no missing answers in the dataset". (Source: ISSP codebooks 2004-2015).

Appendix: Codes for missing values in SDR2

In the SDR database v.2 we identify different situations that warrant to be treated as missing data. Table A.1 lists all SDR2 missing value codes:

Table A.1. Codes for missing values in SDR2

SDR tag [a]	**SPSS (STATA) codes**	**Label**
		Standardized source codes for missing values
DK	-1 (.a)	Don't know
NA	-2 (.b)	No answer
REF	-3 (.c)	Refusal
DU	-4 (.d)	Don't understand the question
DNR	-5 (.e)	Any combination of DK, NA, REF, DU
INAP	-6 (.f)	Inapplicable
NEC	-7 (.g)	Not elsewhere classified
		SDR created codes for missing values
UNFIT	-8 (.h)	Source value does not fit to target
ERR	-9 (.i)	Errors in source data and undocumented source values
COMBI	-10 (.j)	Different missing codes on multiple sources taken for a target
CINAP	-11 (.k)	For control variables only: inapplicable
INSUF	-12 (.l)	For survey: Insufficiently defined response categories
QNA	-20 (.t)	For survey: Question not available

[a] Abbreviations for the labels corresponding to the SDR2 codes for missing values. These tags are used in the Crosswalk Table (CWT) files (Excel) that accompany documentation of SDR2 target variables.

In exceptional situations when codes for missing data listed in Table A.1 cannot be used, we apply a system missing <null> value.