HARMONIZING INFORMATION ON GENDER, YEAR OF BIRTH AND AGE: ATTEMPS TO ELIMINATE INCONSISTENCIES IN 22 INTERNATIONAL SURVEY PROJECTS

Olena Oleksiyenko, Przemek Powałko, Ilona Wysmułek , Anna Franczak

In *Data Harmonization project* we aim to harmonize 1720 surveys conducted in 22 survey projects in 89 waves in 143 countries. The purpose of this report is to focus on differences and similarities in coding of the two basic background variables - age and gender - and describe our approach to the harmonization of these variables.

Surveys for the Harmonia project were chosen based on following criteria:

- Surveys are academic
- Surveys have proper documentation such as questionnaires, codebooks, datasets available in English
- Surveys are cross national covering at least three countries
- Surveys contain questions about protest behavior and democratic values
- Surveys are conducted in at least two waves[1] and on random representative sample of adult population
- Surveys should be accessible free of charge in the public domain

The table below presents lists of survey projects with their full name, acronym, number of waves and data files.

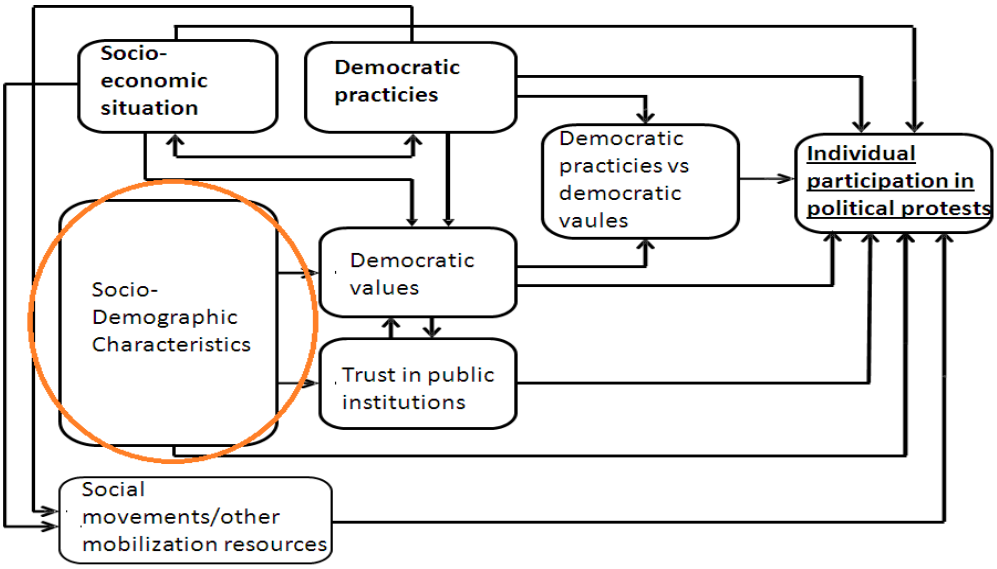
Table 1. List of surveys used in Harmonia project

acronym	survey name	# of waves	Data files
AFB	Afrobarometer	4	AFB/1, AFB/2, AFB/3, AFB/4
AMB	Americas Barometer	5	AMB/1-5
ARB	Arab Barometer	2	ARB/1, ARB/2
ASB	Asian Barometer	3	ASB/1, ASB/2, ASB/3
ASES	Asia Europe Survey	1	ASES
CB	Caucasus Barometer	4	CB/2009, CB/2010, CB/2011, CB/2012
CDCEE	Consolidation of Democracy in Central and Eastern Europe	2	CDCEE/1-2
CNEP	Comparative National Elections Project	1	CNEP/3/ES, CNEP/3/HU, CNEP/3/MX, CNEP/3/MZ, CNEP/3/PT, CNEP/3/TW, CNEP/3/UY, CNEP/3/ZA
EB	Eurobarometer	7	EB/1983, EB/1984, EB/1989, EB/2000, EB/2004, EB/2010, EB/2012
EQLS	European Quality of Life Survey	3	EQLS/1-3
ESS	European Social Survey	6	ESS/1-5, ESS/6
EVS	European Values Study	4	IVS/1-9 (combines dataset for EVS and WVS)

[1] With some exceptions made to expand the time span and coverage of underrepresented countries, for details check Table 1

ISJP	International Social Justice Project	2	ISJP/1-2
ISSP	International Social Survey Programme	13	ISSP/1985, ISSP/1989, ISSP/1990, ISSP/1991, ISSP/1996, ISSP/1998, ISSP/2004, ISSP/2006, ISSP/2007, ISSP/2008, ISSP/2009, ISSP/2010, ISSP/2011
LB	Latinobarometro	15	LB/1995, LB/1996, LB/1997, LB/1998, LB/2000, LB/2001, LB/2002, LB/2003, LB/2004, LB/2005, LB/2006, LB/2007, LB/2008, LB/2009, LB/2010
LITS	Life in Transition Survey	2	LITS/1, LITS/2
NBB	New Baltic Barometer	6	NBB/1-6
PA2	Political Action II	1	PA2
PA8NS	Political Action - An Eight Nation Study	1	PA8NS
PPE7N	Political Participation and Equality in Seven Nations	1	PPE7N_AT, PPE7N_IN, PPE7N_JP, PPE7N_NG, PPE7N_NL, PPE7N_US, PPE7N_YU
VPCPCE	Values and Political Change in Postcommunist Europe	1	VPCPCE_CZ, VPCPCE_HU, VPCPCE_RU, VPCPCE_SK, VPCPCE_UA
WVS	World Values Survey	5	IVS/1-9 (combines dataset for EVS and WVS)

Harmonizing background characteristics is a crucial task to investigate the substantive question the project is about - how democracy impacts soft and hard political protest. The theoretical model explaining protest behavior contains a set of explanatory background variables, such as gender, age, education etc.



Graph 1: Theoretical model explaining protest behavior in the Data Harmonization Project

Background Variables in Survey Research

The variables on age, gender and birth year are the basic background variables used in public opinion survey research. Background variables can be defined as following: "They

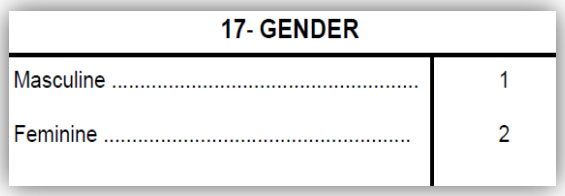
contain information necessary to define homogeneous subgroups, to establish casual relations between attitudes and societal facts, and to define scale scores due to different composition and to identify spurious correlations and casuals relationships. They are used to assess the quality of a realized sample and to decide on any corrections necessary." (Braun and Mohler, 2002: 101, 102).

Gender/sex

Despite the ongoing debate on differences in approach to sex and gender, for the purpose of social science surveys the distinction between male and female individuals is made. There is no clear evidence which would prove that it is better if the question about sex of the respondent is asked by interviewer or interviewer marks the sex of the respondent without asking. However there is a clear recommendation in literature to code Men with 1, Women with 2 (Wolf and Hoffmeyer-Zlotnik 2003: 262).

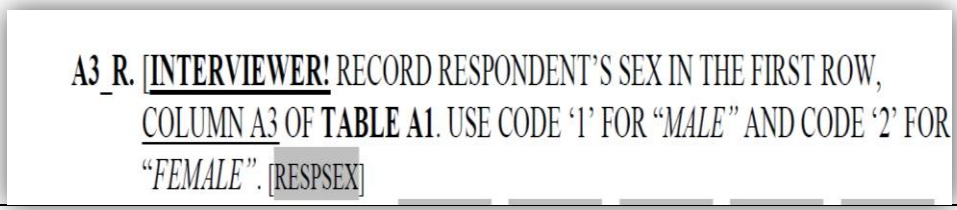
This scheme of sex/gender coding (1-male, 2-female) was used in all 22 survey projects we are examining. Below there are examples of differences in wording used to describe such a universal socio-demographic category. These examples also show that there is no strict rule about whether gender must be coded by interviewer or respondent must be asked about his or her gender.

CNEP/3 Uruguay



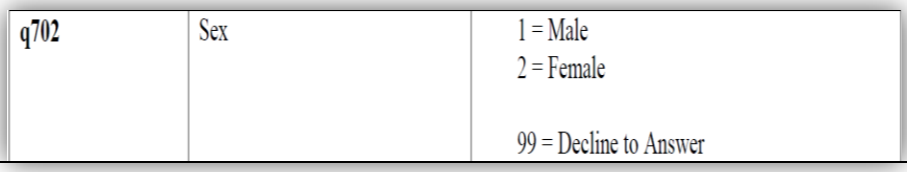
17- GENDER	
Masculine	1
Feminine	2

Caucasus Barometer 2009

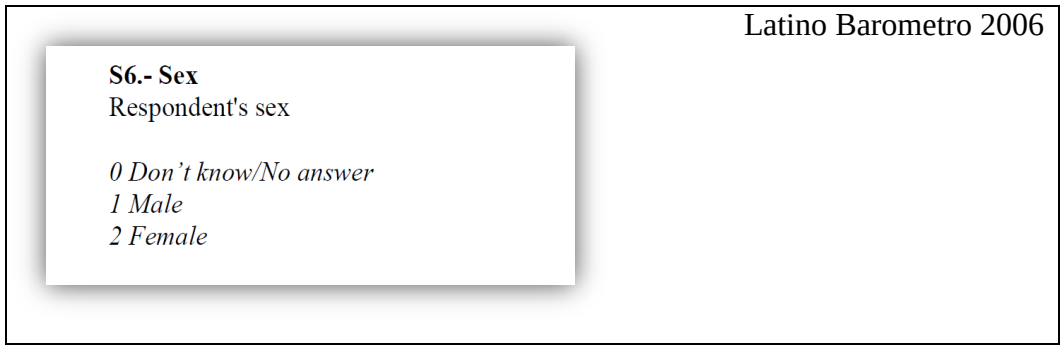


A3_R. [**INTERVIEWER!** RECORD RESPONDENT'S SEX IN THE FIRST ROW, COLUMN A3 OF **TABLE A1**. USE CODE '1' FOR "*MALE*" AND CODE '2' FOR "*FEMALE*". [RESPSEX]

Arab Barometer Wave 1



q702	Sex	1 = Male 2 = Female 99 = Decline to Answer



Graph 2: Differences in Wording. Gender

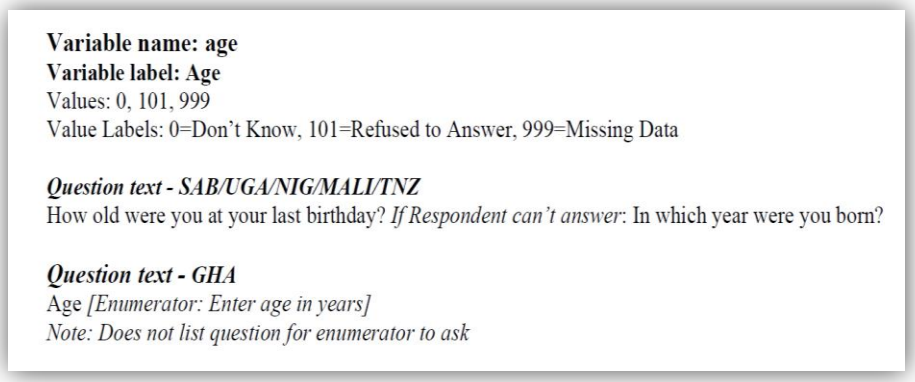
Age

The 'age' variable should not be measured using time intervals, for example 18 to 24, 24 to 34, etc. There are at least two recommendations in literature (Wolf and Hoffmeyer-Zlotnik 2003: 263) on how to measure 'age' of respondent in cross-national surveys. The most recommended is to ask for the date of birth, but in line with the International Standard Date and Time Notation (ISO 8601) YYYY-MM-DD, where YYYY represents the year according to the Gregorian calendar, MM represents the month from January (01) to December (12) and day symbolizes the day of the month from 01 to 31. There are several countries where Gregorian calendar is not used and the method of collecting information on 'age' should be in such case carefully agreed upon.

The second method to assess the 'age' variable with greater accuracy is to simply ask for 'age' directly: "What is your age?" (question asked in line with ESOMAR Standard Demographic Classification). This method is less recommended because of a few reasons: firstly, people more often remember the date of their birth than the accurate age; secondly, the question about age might be perceived as embarrassing while question for date of birth not that much. Moreover, the age limits might be corresponding with legal regulation, for example the year from when voting is allowed. In some countries this is 16 years while in others this is 18 and older. In such case it is important to prepare documentation about it and establish the rules of comparing data (Wolf and Hoffmeyer-Zlotnik 2003: 264).

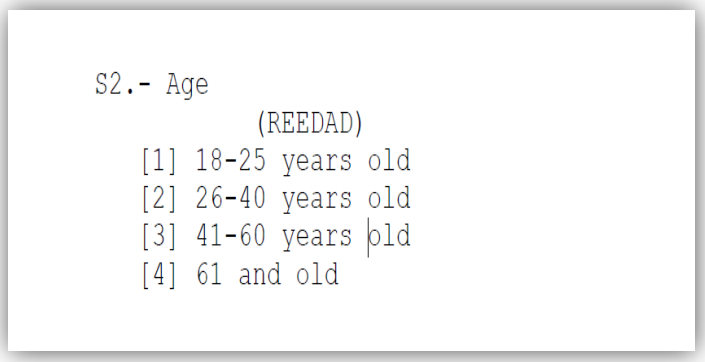
Below there are examples of how age was coded in the surveys included in Data Harmonization Project.

Afro Barometer wave 1



Americas Barometer wave 2

Latino Barometro 2000



Graph 3: Differences in Age Variable coding

Detailed description of background variables in the Harmonia project[2]

Gender

The variable on gender of the respondent, that is a question which was either asked in the questionnaire or coded by an interviewer, appeared in all **22 survey projects**. However, at the level of country/wave/survey, there are five cases in which there is no information about the gender of the respondent:

- Eurobarometer 2012- Montenegro and Serbia
- World Value Survey- Finland, Mexico, and South Africa

[2] Detailed information about each variable described in this report, both on the level of the project-wave and project-wave-country, can be found in the appendix.

It is worth mentioning that the total item non-response for gender is 8062 cases out of 2.3 million. For more detailed information about non-response and missing data on the country and survey level, please, check the Excel supplement file containing this information.

Age

In the project we take into account both information about the age of the respondent (stored in the source variable 's_age') and the birth year of the respondent ('s_birth_year'). Moreover, in order to recalculate the age from the year of birth or to make more accurate estimates, we collect the information about the date of conducting surveys ('s_interview_date'). This information is often available in the dataset or documentation.

The variable on age of the respondent appeared in **all waves of 22 survey projects**. However, there are waves in which question on age was not asked in some countries. These are the following cases:

- in Americas Barometer, wave 3 there is no information about the age of respondent for the United States of America;
- in the first wave of Consolidation of Democracy in Central and Eastern Europe- no information on age in Lithuania and Romania;
- in the first wave of World Value Survey- no age question in Finland, Hungary, South Korea, Mexico and South Africa. There is also an interesting case of Argentina and Japan in the same survey-wave where only some random respondents report their age.

The information about the birth year of respondent ('s_birth_year') appeared in **30 waves of 12 survey projects**.

On the level of country, there is no information about birth year in:

- Comparative National Election Project for Spain, Hungary, Mozambique, Portugal, and South Africa;
- Political Participation in Seven Nations- Austria, India, Japan, Nigeria, United States of America, and Yugoslavia.

There are at least two exceptional cases, where questions were asked about the birth year, but in data there is obviously exact age derived from the birth year coded (CNEP/3 Uruguay and PPE7N The Netherlands). Age is also derived from birth month, birth year and date of interview in International Social Justice Project 1-2 and Political Action 2, so we use age variable for these two surveys.

Despite the recommendations not to code age in brackets, there are several exceptions from this rule we encountered:

1. In case of Political Participation and Equality in Seven Nations, we have age coded in brackets for three countries: India, Japan and Yugoslavia.

2. In Arab Barometer wave 1 there is no exact age for Morocco coded and we use variable with age coded for all countries in brackets to create the coherent variable for this survey.

3. Similar case is in Asian Barometer wave 1: we have two variables about respondent's age - exact age in years and age in brackets, but for the first one information about Mongolia is missing.

4. In International Social Survey Project 1985, all countries report information about exact age, but in case of Italy age in coded in brackets.

5. In Caucasus Barometer 2009 information about how the exact age of respondent was calculated is missing, but we have both the year of birth and age brackets for this survey.

In **46 waves of 15 survey projects** we have the information about **the interview year**. However, there are cases in which interview year is different in the documentation and in the data set. We treat the interview year from the data set as the more accurate information. Moreover, the additional research of documentation was made in order to extract the exact information about the year of conducting survey for each country/survey/wave unit. The years assigned for each unit are coded as the target variable 't_country_year'**.** The target variable contains information about exact year of conducting the fieldwork, if it appeared in the data set. In cases when there was no information in the data set, the year from the country-specific codebooks was assigned. If in the codebook there was not one exact year for a country, but the time span of couple years:

a) the information on the number of months was taken into account and the year in which there was the greater number of months of the fieldwork done was chosen for the target variable;

b) the year with greater number of respondents surveyed was treated as the target variable number.

In project/waves that have a specific variable containing the year of conducting the survey, for some countries this information is still missing. The examples of such surveys are below. On the country level, information about interview data is missing for:

- Asian Barometer wave 2- South Korea, Philippines, Singapore, and Thailand;
- Comparative National Election Project- Spain, Mexico, Mozambique, Portugal, Taiwan, Uruguay, and South Africa.

Procedures of standardization across 22 international survey projects

Gender

Since all the variables were coded according to the same scheme, we did not use any additional harmonization procedures.

Age and Year of Birth

- We decided that values from 14 to 96 are included to the scope of age as a target variable. This decision is based on two assumptions: 14 is the lowest value for age documented in survey documentation and 97, 98 and 99 are often codes for missing data. To avoid mixing missing data with real age we decided to put the upper limit at 96. Age above 100 years is always coded as missing data. Age 97, 98, 99 is coded either as missing data (if documented) or 96+ if indicates real age.

- If the question about year of birth was asked, we calculate age based on this variable and the interview date. Priority is always given to the variable constructed in this way, even if a source variable[3] describing the age is available. (Situations when we have both year of birth and exact age of respondent are rather uncommon. This is the case of Asian Barometer wave 3, all four waves of Caucasus Barometer, Consolidation of Democracy in Central and Easter Europe wave 1 and 2, Comparative National Election Project Mexico, European Social Survey waves 1-6, all waves of World Value Survey and European Values Survey, International Social Justice Project 1-2, International Social Survey Program 2010 and 2011, Political Action - An Eight Nation Study, Political Action II.)
- If there is no information about the year of birth, we use the age variable from the source data.
- In case we have respective source variables but the values are spurious, we try to correct them. For example, in World Value Survey wave 1 we found an invalid and undocumented value for the year of birth, but in the same time the value of age variable is valid, so we use the age variable in these cases while in all other we still use the year of birth for calculating the age.
- If the age or the birth year variables cannot be corrected, they are marked as missing data (e.g. the age below 10 in Latino Barometro 1995-1997, empty values in Asian Barometer wave 3 etc.). We always look for additional information about suspicious/undocumented values in order to correct them.
- If age is in brackets, we take the midpoints (29=25-34) or edge points (75=75+,19=19-) (Arab Barometer wave 1, Asian Barometer wave 1, Caucasus Barometer 2009 International Social Survey Program 1985, PPE7N-Iidia, Japan and Yugoslavia)

Interview year

- If the variable about the exact interview date is not available, the year of fieldwork is taken from the survey documentation. Priority is the date of survey found in the data set.
- If the fieldwork lasted for more than one year, we choose the year with the majority of respondents interviewed or with the longer period of fieldwork conducting.
- Sometimes there is no information about fieldwork dates even in survey documentation and we had to guess the year of survey from other sources or to arbitrarily choose the exact year from possible values. For example, there is no sufficient information about Comparative National Election Project wave 3, Mozambique. From the project's webpage we know that release date was 2005, but is it a year of the fieldwork, since the survey dealt with elections held in the beginning of December 2004.

Obvious errors

Gender

The only problematic question about respondent's gender was encountered in Life in Transition wave 2. Variable **respondentgender** appeared only in the data and there are no

[3] Original variable before harmonization

labels for variable values 1 and 2. In case of other surveys we either had unexplained missing data or problems with translation from Spanish to English in Americas Barometer 1-5.

Age

- Age- above 97 years and below 15 (or depending on the definition of target population) if not documented that it is minimum and maximum age of respondent

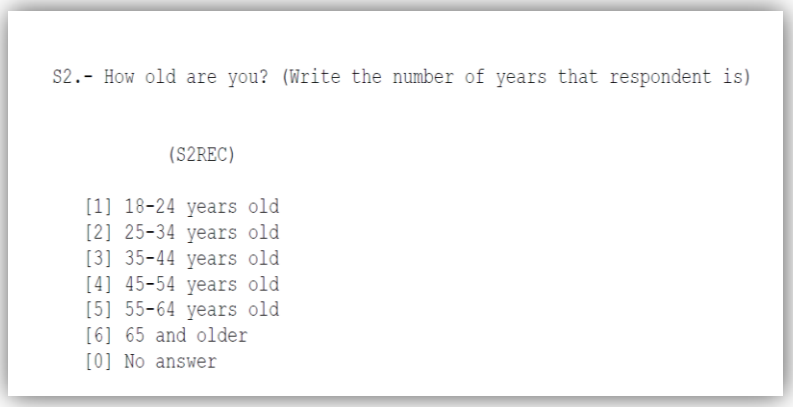
This error appeared in almost every survey analyzed, the only exceptions were European Quality of Life Survey and Asia Europe Survey.

- In Consolidation of Democracy in Central and Eastern Europe 1 we have over-representation of the value 99 compared to the rest of 90's while the dictionary codes NA as 999, the value which happens in dataset along with 99
- In Latino Barometro (all waves) we also encountered the problem of misleading presentation of age categories

Example:

Latino Barometro 1997

Exact question wording:



In data, we have values indicating the exact age of respondent

null 0 2 7 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 97 99

In country-specific codebooks available in English, the age brackets for the same variable are different: 15-25,26-40,41-60, 60 and more. This inconsistency is repeated from wave to wave.

Year of birth

- Year of birth is approximate – for example: in Caucasus Barometer 2009,2010 some respondents do not report their age and don't remember exact year of birth, so we end up with values given in decades: 1890s, 1900s, 1910s etc.

Interview year

- Year of survey is incorrect – for example: in Caucasus Barometer 2010 we expect 2010 but sometimes we found 1959 in the dataset.

Summary

This paper aims to summarize the decisions made by the team concerning harmonization of two basic socio-demographic characteristics: age (age in years and year of birth) and gender - in cross-national surveys included in Data Harmonization project. We presented the differences in coding of these variables, problems encountered and implied solutions. We showed the variety of approaches used by investigators around the world and lack of unified standards envisioned in most of methodological recommendations, which were also briefly discussed. In reports on data quality most of this problems are described in detailed manner.

REFERENCES:

1. Braun, M. and Mohler, P. Ph. (2002). Background Variables. In: A. Harkness, F.J.R. Van de Vijver and P.Ph. Mohler (eds.), *Cross-Cultural Survey Methods*. New York: Wiley.
2. Harkness, J.A., Hoffmeyer-Zlotnik, J.H.P., (2005) *Methodological aspects in Cross-National Research*. ZUMA.
3. Hoffmeyer-Zlotnik, J.H.P., (2008) *Harmonisation of Demographic and Socio-Economic Variables in Cross-National Survey Research. Bulletin of sociological methodology*.
4. Kolsrud, K., and Kalgraff Skjåk, K. (2005). Harmonising background variables in the European Social Survey. In J.H.P. Hoffmeyer-Zlotnik and J.A. Harkness (Eds.), *Methodological aspects in cross-national research.* ZUMA.
5. Wolf, C., and Hoffmeyer-Zlotnik, J.H.P. (2003). Measuring Demographic and Socio-Economic Variables in Cross-National Research: An Overview. In: J.H.P. Hoffmeyer-Zlotnik and C. Wolf (eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*. New York: Kluwer Academic/Plenum Publishers
6. Wolf, C., and Hoffmeyer-Zlotnik, J.H.P. (2003). How to measure e/Gender and Age In: J.H.P. Hoffmeyer-Zlotnik and C. Wolf (eds.), *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*. New York: Kluwer Academic/Plenum Publishers