Identification of Processing Errors in Cross-national Surveys

Olena Oleksiyenko, Ilona Wysmulek, Anastas Vangeli

Institute of Philosophy and Sociology Polish Academy of Sciences,
Graduate School for Social Research

3MC International Conference
Chicago
July 25 - 29, 2016

Research supported by the Polish National Science Centre (2012/06/M/HS6/00322) as part of the study *Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling*

Theoretical Considerations

- Processing error refers to the “bias and variance that may result from mistakes made while processing data, including the coding and recoding of data, the transformation of data into new variables, the imputation of missing data, the weighting of the data, and the analyses that are performed with data” (Lavrakas 2008, p. 539)
- (In)consistencies between survey documentation and records in the data are a specific type of processing error
- The concepts and tools for analyzing processing errors are underdeveloped and limited in number

Framework for Identification of Process Errors

- Starting point: the need to establish quality control procedures
- Focus: the consistency between different sources of documentation and numeric data records
- Data: seven target variables of substantive interest for the Harmonization Project; a total of 688 source-variables from the 89 survey waves
- Procedure: cross-checking documentation-data files
- Outcome: 153 processing errors identified; creation of typology of errors

Analytical Steps

- Unit of analysis - variable on the survey wave level
- **7 target variables (5 demographic + 2 substantive)**; a total of **688** source variables correspond to them
- For each source variable, we analyzed (a) codebook and/or questionnaire; (b) SPSS dictionary and (c) exact records of data from computer files regarding the following aspects:

1. variable name (codebook/questionnaire);
2. exact questionnaire-item formulation (codebook/questionnaire);
3. variable label (codebook/SPSS dictionary)
4. value labels (codebook/questionnaire/SPSS dictionary)
5. numerical values (codebook/questionnaire/SPSS dictionary/data records)

Sample

#	**Target Variables**	**# Source var. /wave**		**# Waves**
		Min	**Max**	
1	Gender	1	1	89
2	Age	1	3	89
3	Birth Year	0	1	29
4	Education levels	0	18	76
5	Schooling years	0	2	72
6	Trust in parliament	0	1	67
7	Participation in demonstration	0	4	65

*all sample: **688** source variables matched to target variables

Methodological challenges: How to analyze the relationship between metadata and numerical data?

Method: **content analysis and coding**

Five stages:

1. Data collection
2. Pre-coding
3. Formulation of categories
4. Assigning categories
5. Reliability check

Results: types of processing errors

1. Illegitimate Variable Values
2. Misleading Variable Values
3. Contradictory Variable Value
4. Variable Values Discrepancy
5. Lack of Labels for Missing Data

Example:

Survey code	Variable name	Question wording	Values from codebook and SPSS dictionary	Data	Target Variable
EQLS/1	Y07_CVq48	How old were you when you completed your full-time education?	96 = Still in education 97 = Never completed 98 = Don't know 99 = Refusal - only separated in 2003	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 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 96 97 98 99 997 998 999	Schooling years

Results: types of processing errors

1. Illegitimate Variable Values
2. Misleading Variable Values
3. Contradictory Variable Value
4. Variable Values Discrepancy
5. Lack of Labels for Missing Data

Example:

Survey code	Variable name	Question wording	Values from codebook	Values from SPSS dictionary	Data	Target Variable
LB/1995	s2	How old are you? (Write the number of years that respondent is)	[1] 18-24 years old [2] 25-34 years old [3] 35-44 years old [4] 45-54 years old [5] 55-64 years old [6] 65 and older	1|18-24 2|25-34 3|35-44 4|45-54 5|55-64 6|65 y +	0 5 8 15 16 17 18 19 20 21 22 23 24 25 […] 84 85 87 88 89 90 93 94 96 99	Age

Results: types of processing errors

1. Illegitimate Variable Values
2. Misleading Variable Values
3. Contradictory Variable Value
4. Variable Values Discrepancy
5. Lack of Labels for Missing Data

Example:

Survey code	Variable name	Question wording	Values from codebook	Values from SPSS dictionary	Data	Target Variable
LITS/2	q303e	To what extent do you trust the following institutions? /people. The parliament	1 Complete distrust 2 Some distrust 3 Neither trust nor distrust 4 Some trust 5 Complete trust 6 Not applicable 7 Don't Know	-99Refused -98Not applicable -97Don't know -90Filtered -1 Not stated 1Complete distrust 2Some distrust 3 Neither trust nor distrust 4 Some trust 5 Complete trust 6 Not applicable	-98 -97 - 1 1 2 3 4 5	Trust in Parliament

Results: types of processing errors

1. Illegitimate Variable Values
2. Misleading Variable Values
3. Contradictory Variable Value
4. Variable Values Discrepancy
5. Lack of Labels for Missing Data

Example:

Survey code	Variable name	Question wording	Values from codebook	Values from SPSS dictionary	Data	Target Variable
ASB/2	q010	I'm going to name a number of institutions. For each one, please tell me how much trust do you have in them? Parliament	1. A great deal of trust 2. Quite a lot of trust 3. Not very much trust 4. None at all 7.DU 8.CC 9.DA	1.None at all 2.Not Very Much Trust 3.Quite a Lot of Trust 4.A Great Deal of Trust 7.Do not understand the question 8.Can't choose 9.Decline to answer	null 1 2 3 4 7 8 9	Trust in parliament

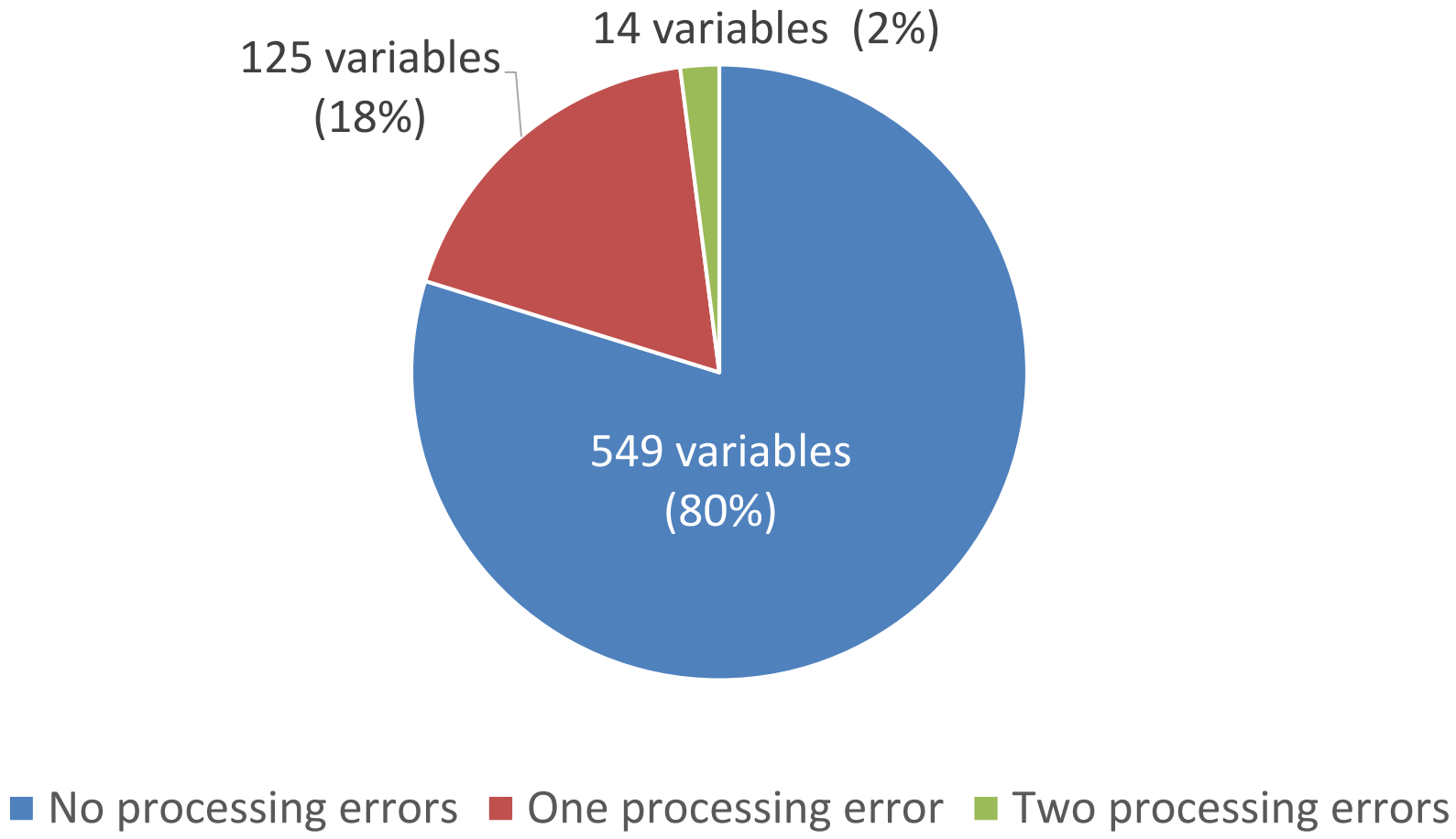
Results: types of processing errors

1. Illegitimate Variable Values
2. Misleading Variable Values
3. Contradictory Variable Value
4. Variable Values Discrepancy
5. Lack of Labels for Missing Data

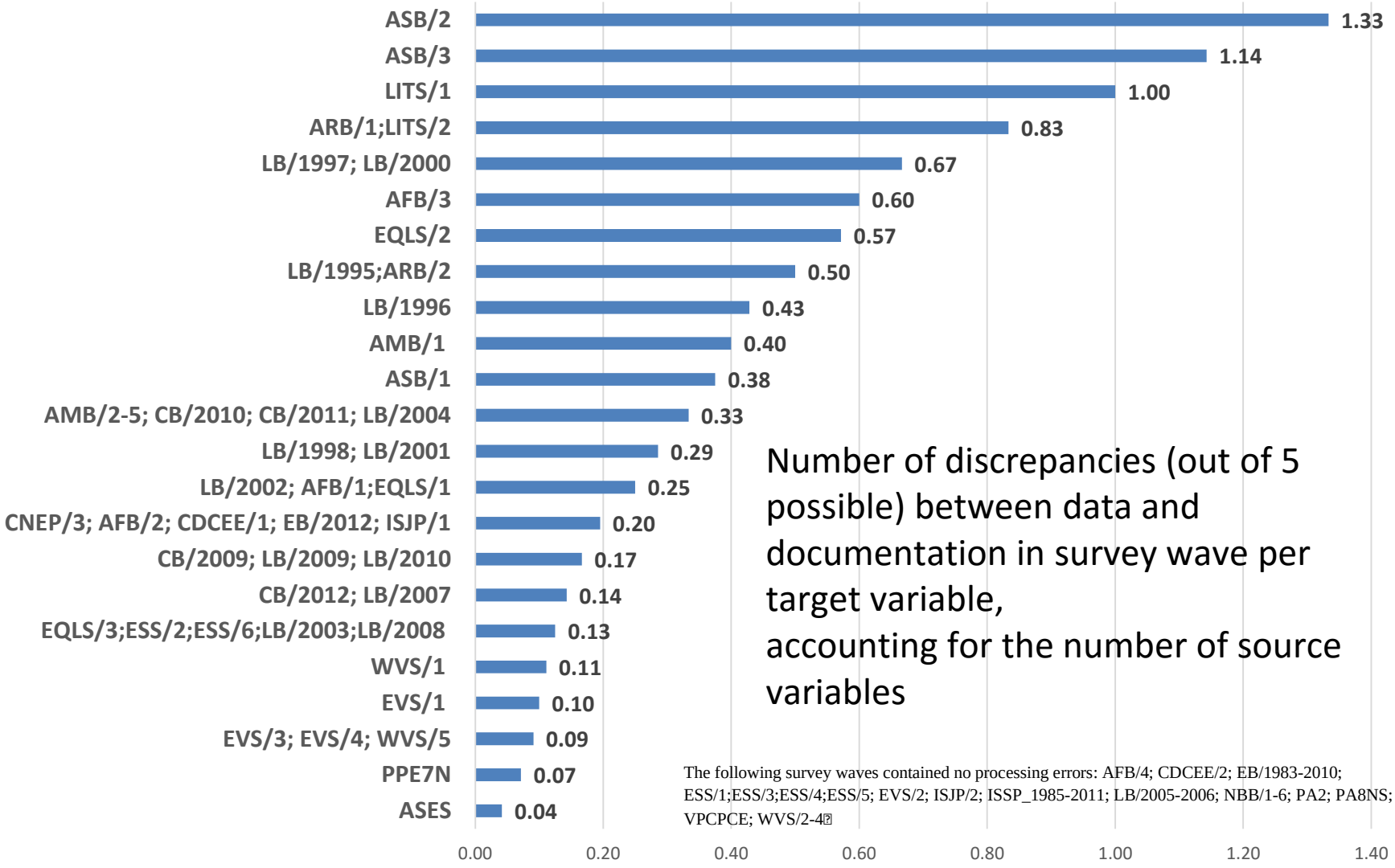
Example:

Survey code	Variable name	Question wording	Values from codebook	Values from SPSS dictionary	Data	Target Variable
WVS/1	X003R2	This means you are __ __ years old.	-5 other missing -4 question not asked -3 not applicable -2 no answer -1 don't know 1 15-29 years 2 30-49 years 3 50 and more years	.-5 Missing; Unknown -4 Not asked in survey -3 Not applicable -2 No answer -1 Don´t know 115-29 years 2 30-49 years 3 50 and more years	.-5 -4 1 2 3 14	Age

Prevalence


14 variables (2%)
125 variables (18%)
549 variables (80%)
No processing errors
One processing error
Two processing errors

Distribution of errors per project-wave (unweighted index)



Number of discrepancies (out of 5 possible) between data and documentation in survey wave per target variable,
accounting for the number of source variables

The following survey waves contained no processing errors: AFB/4; CDCEE/2; EB/1983-2010; ESS/1;ESS/3;ESS/4;ESS/5; EVS/2; ISJP/2; ISSP_1985-2011; LB/2005-2006; NBB/1-6; PA2; PA8NS; VPCPCE; WVS/2-4

Conclusion

- Data-processing errors render the documentation less ‘user friendly’ and lower the fitness of the data for use;
- At the stage of data production and management, the typology of processing errors can be used as a check list to strengthen total survey quality
- At the stage of ex-post survey data harmonization one can construct indicators of specific error types and use them as controls in analysis
- Aspects to be considered in the future: applying “severity weights” and possibily automatizing the process