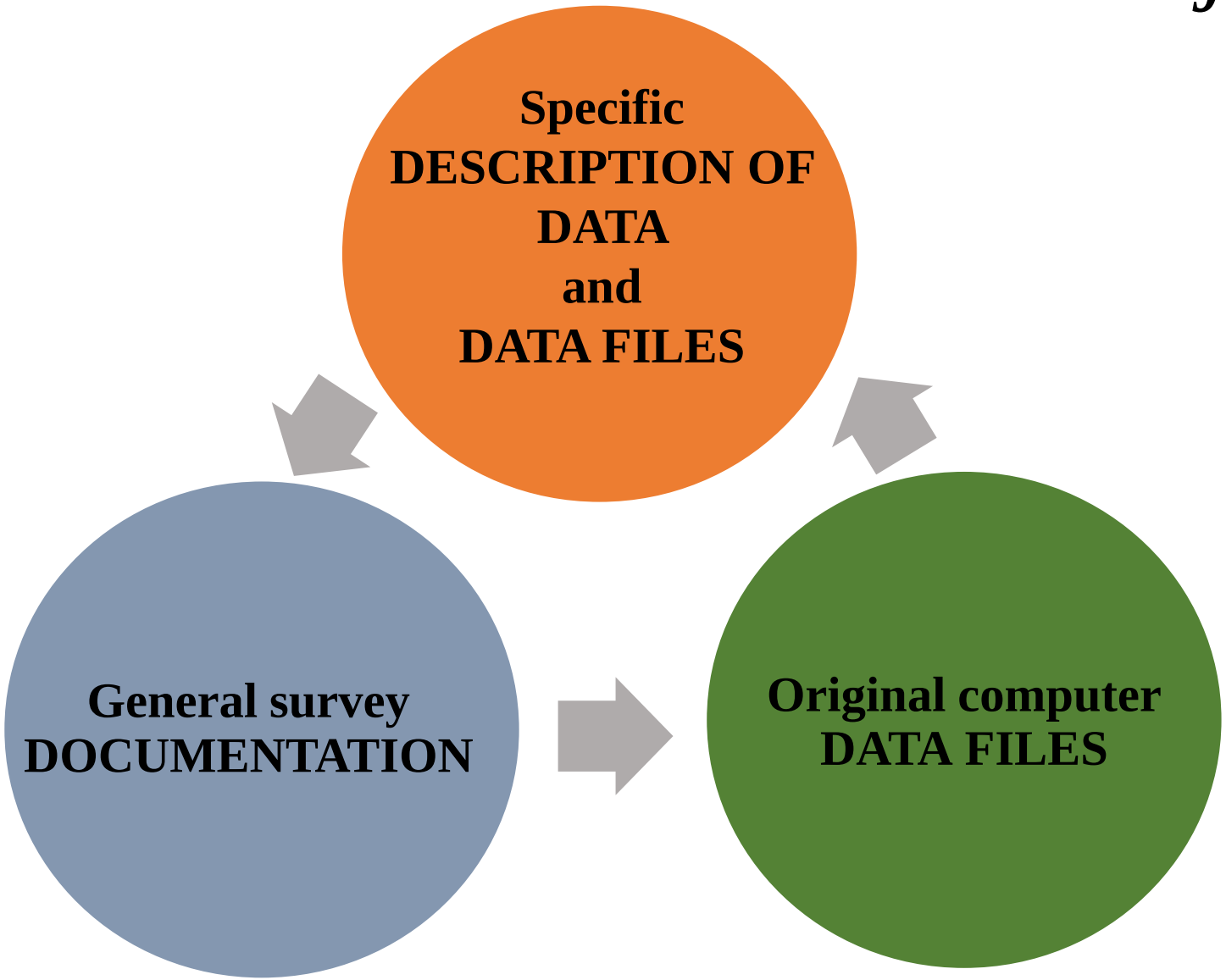
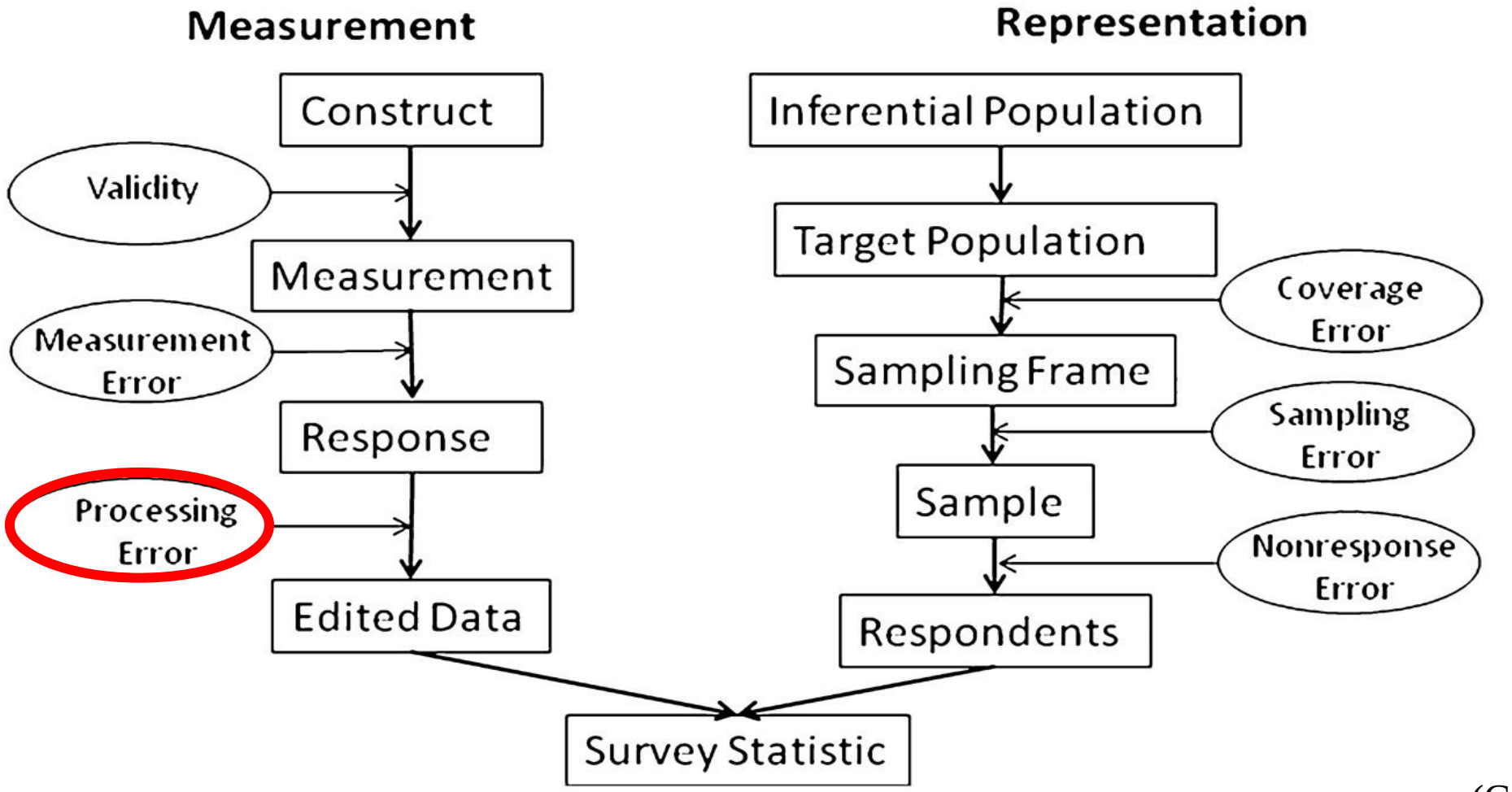
(In)Consistency of Survey Data and their Description

Quality Realms in the *Data Harmonization Project*



Total Survey Error



(Groves 1989)

Working with both: data AND documentation

- “coding, editing, … and other data processing activities that follow the data collection phase” might be the source of errors (Groves 1989:12). [recognized in the literature]
- But… „[processing errors are] too rarely included in models of survey error” (Groves 2010: 869)

Sample

#	**Target Variables**	**# Source var. /wave**		**# Waves**
		Min	**Max**	
1	Gender	1	1	89
2	Age	1	3	89
3	Birth Year	0	1	30
4	Education levels	0	18	76
5	Schooling years	0	2	75
6	Trust in parliament	0	1	77
7	Participation in demonstration	0	4	75

Sample: **687** source variables matched to target variables

Sources

Data gathered from:

a) codebook and/or questionnaire
b) SPSS dictionary and original ‚raw' data

Basic information: a) variable name b) exact question formulation c) variable label (codebook/SPSS dictionary) d) value labels (codebook/questionnaire/SPSS dictionary) e) exact values in the data

Acknowledgements ☺

Anna Franczak
Anastas Vangeli
Jakub Wysmułek

Data&Documentation Discrepancy - Types

VARIABLE VALUE	VARIABLE LABEL	INFORMATION
1. Illegitimate Value (IV)	6. Misleading Label (ML)	7. Insufficient Information (II)
2. Misleading Value (MV)		8. Translation Issues (TI)
3. Value Discrepancy (VVD)		
4. Contradictory Values (CVL)		
5. Lack of Labels (LL)		

Illegitimate Value
Misleading Value

Survey code	Variable name in the data set	Question/ Questionnaire item	Label in codebook	Label in SPSS dictionary	Values from codebook/ questionnaire	Values from SPSS dict	Data	Target variable
LB/1995	s2	How old are you? (Write the number of years that respondent is)	-	EDAD ENTREVISTADO	[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 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 87 88 89 90 93 94 96 99	Age

Value Discrepancy

Survey code	Variable name	Question/ Questionnaire item	Label in SPSS dict	Values from codebook/ questionnaire	Values from SPSS dict	Data	Target variable
CNEP/3/ ZA	H.Protest	Have you taken part in a protest demonstration in the last twelve months?	Q47. Participation in Protest Demonstrations	0-No 1-Yes Don’t know [Do not read] 99	1.00|No 2.00|Yes 3.00|Don't know (Do not read out)	1 2 3	PR_demonst

Contradictory Values

Survey code	Variable name	Question/ Questionnaire item	Label in SPSS dict	Values from codebook/ questionnaire	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	Trust in 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	Tr_parli

Lack of Labels [for Missing Data]

Survey code	Variable name in the data set	Question/ Questionnaire item	Label in SPSS dictionary	Values from codebook	Values from SPSS dictionary	Data	Target variable
PPE7N_NL	V345	Have you ever taken part in a demonstration or protest march?	PROTEST PARTIC IN DE	(1) Yes (5) No (8) Don't know (9) NA	-	0 1 3 5 8	PR_demonst

Misleading Variable Label

Survey code	Variable name	Question/ Questionnaire item	Label in codebook	Label in SPSS dict	Values from codebook	Values from SPSS dict	Data	Target variable
ARB/1	q2302	Here is a list of actions that people sometimes take as citizens. For each of these please tell me whether you, personally, have ever done each of these things in the past three years. Attend a demonstration or protest march	Attend a demonstration or protest march	q230.2. here is a list of actions that people sometimes take as citizens. for ea	1 = Once 2 = More than once 3 = Never 97 = Not clear 98 = Can’t Choose/Don’t know 99 = Decline to Answer	1|once 2|more than once 3|never 97|not clear 98|can't choose/don't know 99|decline to answer	null 1 2 3 97 98 99	PR_demonst

Insufficient Information

Survey code	Variable name in the data set	Question/ Questionnaire item	Label in codebook	Label in SPSS dictionary	Values from codebook/questionnaire	Values from SPSS dictionary	Data	Target variable in Harmonia
LITS/2	respondentgender	~~	~~	~~	~~	~~	-1 1 2	Gender

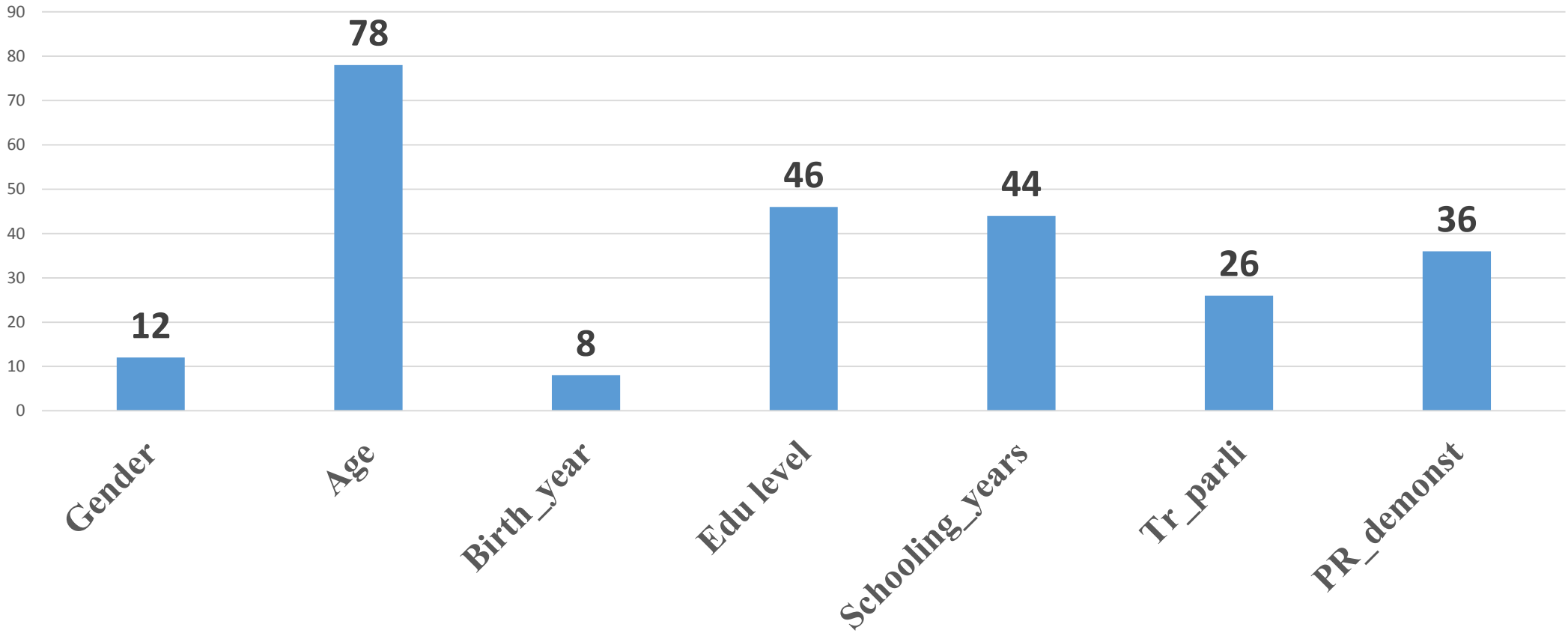
Translation Issues

Survey code	Variable name in the data set	Question/ Questionnaire item	Label in SPSS dictionary	Values from codebook/questionnaire	Values from SPSS dictionary	Data	Target variable in Harmonia
AMB/1	prot1	Have you ever participated in a public demonstration or protest? Do you do it often, rarely or never?	Alguna vez en su vida, ¿ha participado usted en una manifestación o protesta púb	(1)Sometimes (2)Rarely (3)Never (8)DN	1|Algunas veces 2|Casi nunca 3|Nunca 888888|DK 988888|NR 999999|N/A	1 2 3 888888 999999	PR_demonst

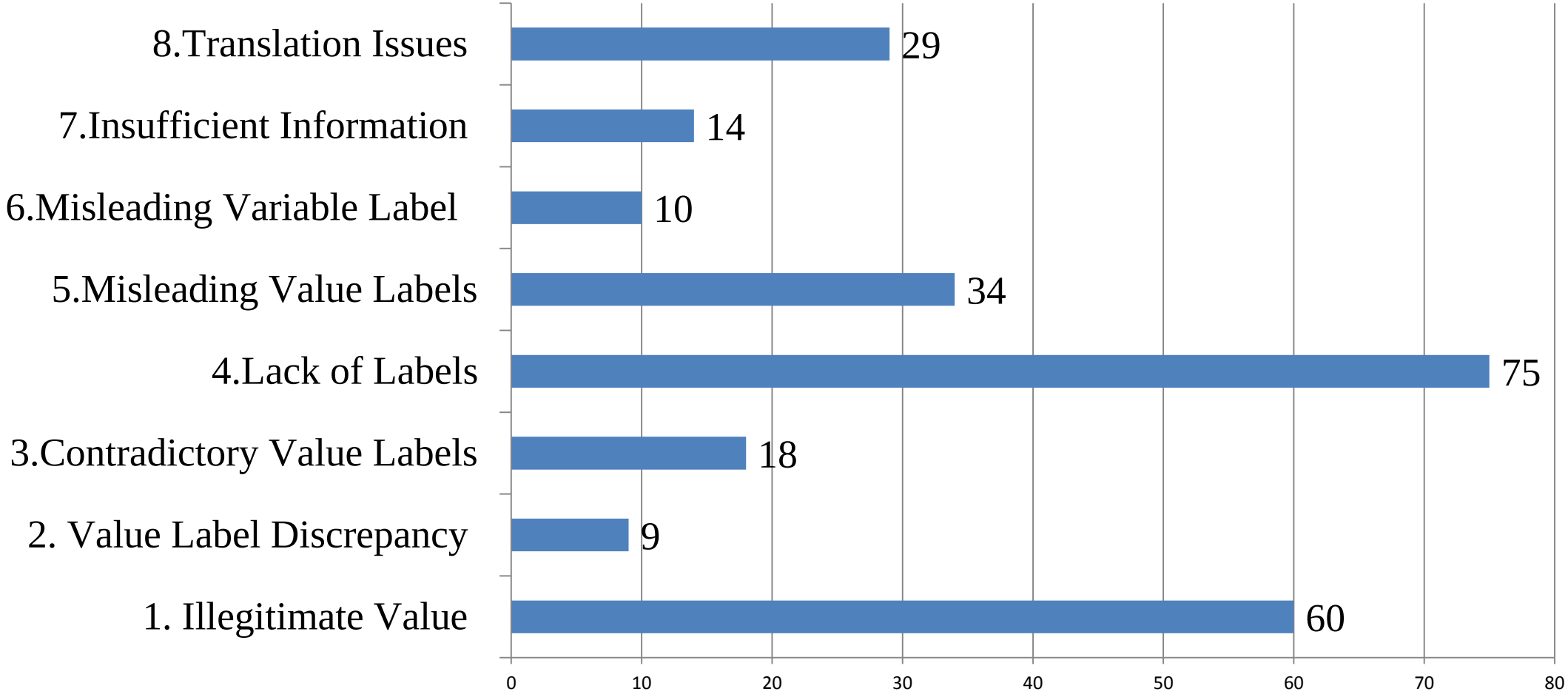
Frequencies of Discrepancies [total: 687 source variables]

		Frequency
Total number of variable containing errors		**197**
Number of errors	1	146
	2	49
	3	2
Total number of errors		**250**

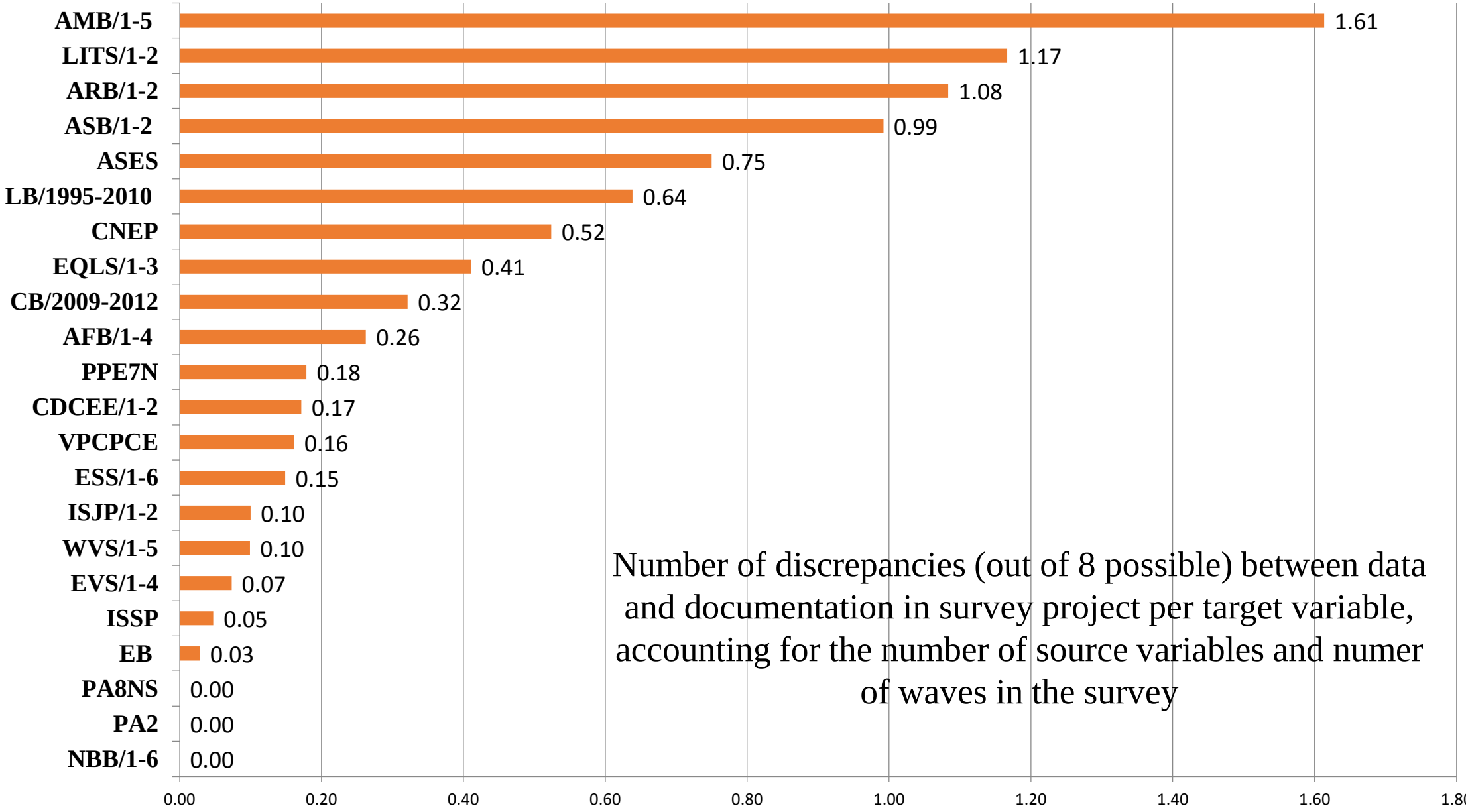
Discrepancies per Target Variables [total =250 discrepencies]



Types of errors
[total =250 discrepencies]



QUALITY INDEX per Survey Project



Number of discrepancies (out of 8 possible) between data and documentation in survey project per target variable, accounting for the number of source variables and numer of waves in the survey

Pearson Correlation

	1.IV	2.VLD	3.CVL	4.LL	5.MVL	6.MVrL	7.II	8.TI
1. Illegitimate Value (IV)	1	0,092	0,212	0,022	0,354	0,133	0,186	0,128
2. Value Label Discrepancy (VLD)		1	0,121	0,463	0,007	0,121	0,34	-0,05
3.Contradictory Value Labels (CVL)			1	0,057	-0,082	-0,099	0,054	0,776
4.Lack of Labels (LL)				1	-0,053	0,16	0,189	-0,11
5.Misleading Value Labels (MVL)					1	0,288	0,329	-0,14
6.Misleading Variable Label (MVrL)						1	0,054	-0,08
7.Insufficient Information (II)							1	-0,07
8.Translation Issues (TI)								1

CDCEE/1-2 AGE V580

V580 AGE (IN YEARS)

Decimals: 0
Missing values: 0 ; 999 ;

Question: Age in years.

Value	Label
0	NAV (Other c.)
999	NA

CODEBOOK

```
CDCEE_1_2 v580
Age (in years)
0|NAV (Other c.)
999|NA

in dataset: 0 12 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 94 98 99 999
```

SPSS DICTIONARY & DATA

CNEP/4, L.EDUCATION

67 What is the highest level of education you have completed? *[Do not read options. Code from answer.] L.Education*	
No formal schooling (cannot read or write)	0
Some primary schooling	1
Primary school completed	2
Some secondary school / high school	3
Secondary school completed (vocational or commercial school)	4
Secondary school completed / high school (general educational track)	5
Incomplete university education / Other post-secondary qualifications (e.g. diploma / degree from a technikon or college)	6
University completed	7
Post-graduate	8
Don't know *[Do not read]*	99

CODEBOOK

```
CNEP_3_ZA L.Education
Q67. HIGHEST LEVEL OF EDUCATION
1.00|No formal schooling (cannot read or write)
2.00|Some primary schooling
3.00|Primary school completed
4.00|Some secondary school / high school
5.00|Secondary school completed (vocational or commercial school)
6.00|*Secondary school completed / high school (general education
7.00|Incomplete university education /Other post-secondary qualif
8.00|University completed
9.00|Post-graduate
10.00|Don't know [Do not read]

in dataset: 1 2 3 4 5 6 7 8 9
```

SPSS DICTIONARY & DATA

Discussion

- Discrepancies between data and documentation = decrease of interpretability of data
- How discrepancy typology can be used?
- Potential quality threat: different weight of types of discrepancies

- 20% of discrepancies out of all variables checked – is it a lot?