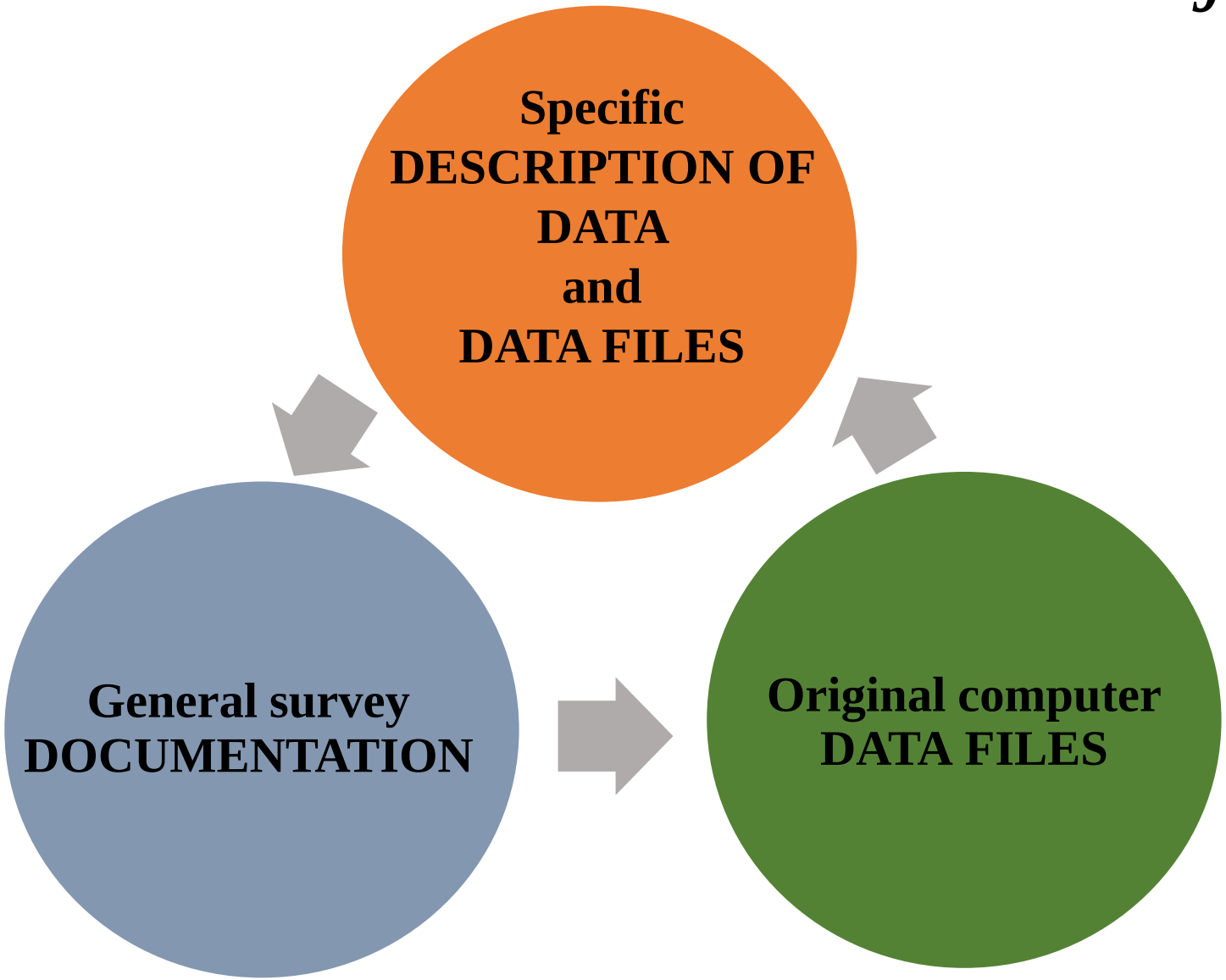
2015 CSDI Workshop, March 26th-28th at City University in London

# Documentation Quality Assessment in Ex-Post Survey Harmonization: Implications for Comparative Research

**Ilona Wysmułek**
*ilona.wysmulek@ifispan.waw.pl*
**Olena Oleksiyenko**
*olena.oleksiyenko@gmail.com*
**Polish Academy of Science, Institute of Philosophy and Sociology**

# Quality Realms in the *Data Harmonization Project*



# Working with: data AND documentation

- Interpretability of the data: does good documentation mean good quality of data?
- “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

# Data and Documentation Discrepancies: 8 types

| VARIABLE LABEL | VARIABLE VALUE | INFORMATION |
|---|---|---|
| 1.Misleading Label | 2. Illegitimate Value | 7.Insufficient Information |
| | 3.Misleading Value | 8.Translation Issues |
| | 4.Value Discrepancy | |
| | 5.Contradictory Values | |
| | 6.Lack of Labels | |

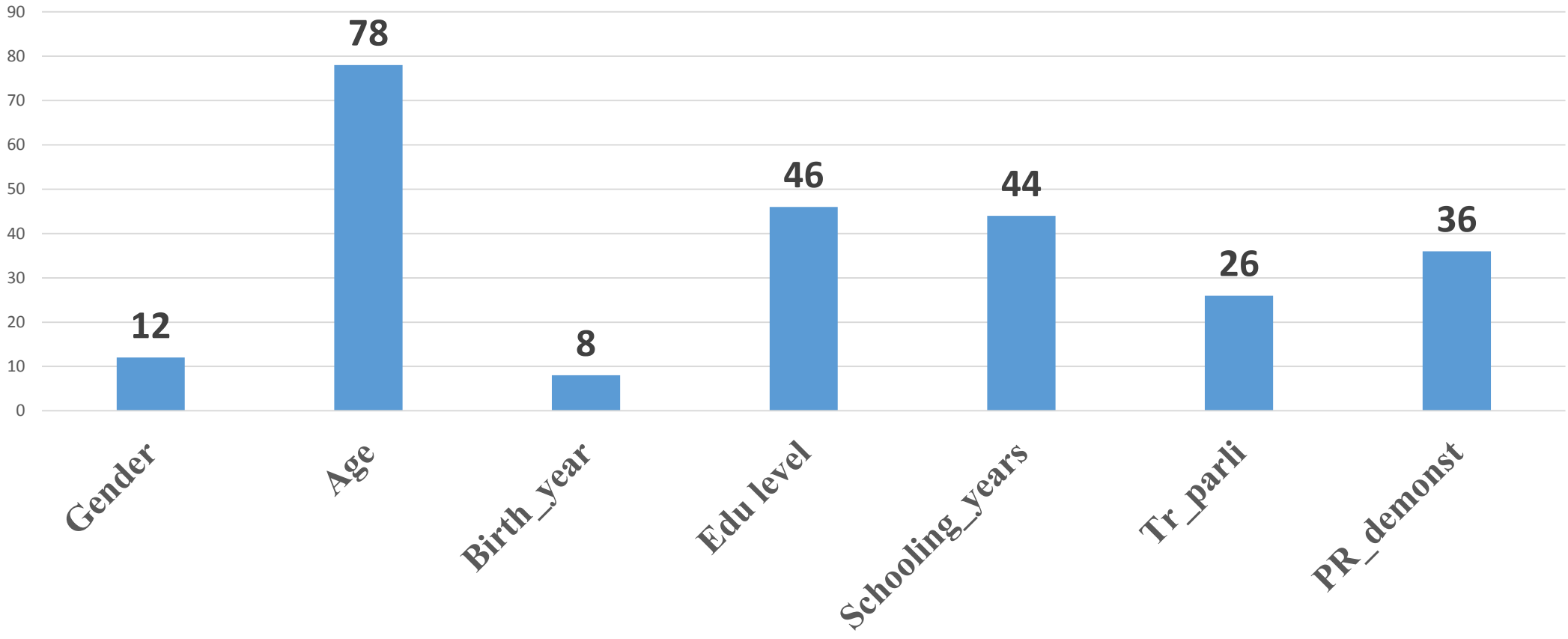
Example:

| Survey | Question | Codebook | SPSS dictionary | DATA |
|---|---|---|---|---|
| ASB/2 | Have you attended a demonstration or protest march? | Once 1<br>More than once 2<br>Never Done 3<br>CC 8<br>DA 9 | 1\|Once<br>2\|More than once<br>3\|Never<br>8\|Can't choose<br>9\|Decline to answer | null 1 2 3 7 8 9 |

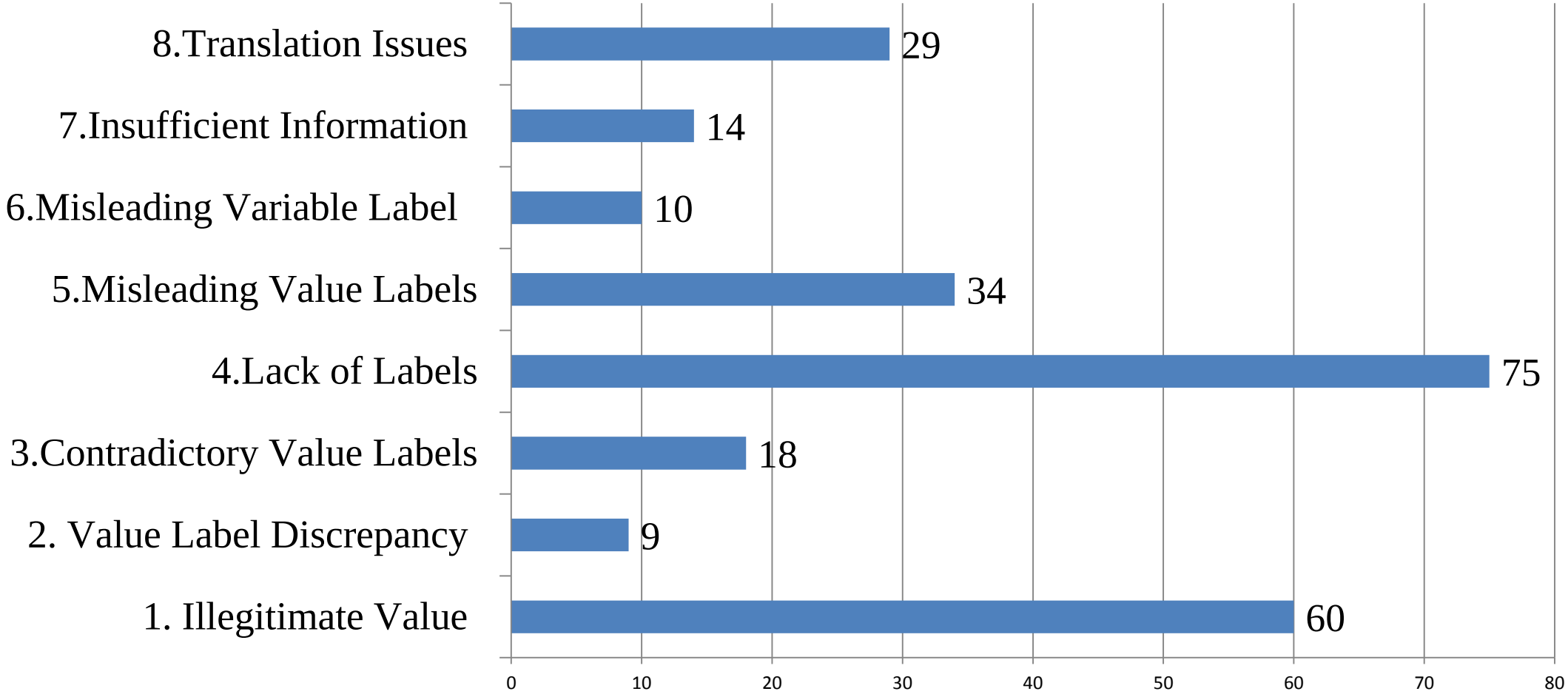
# Frequencies of Discrepancies [total: 687 source variables]

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

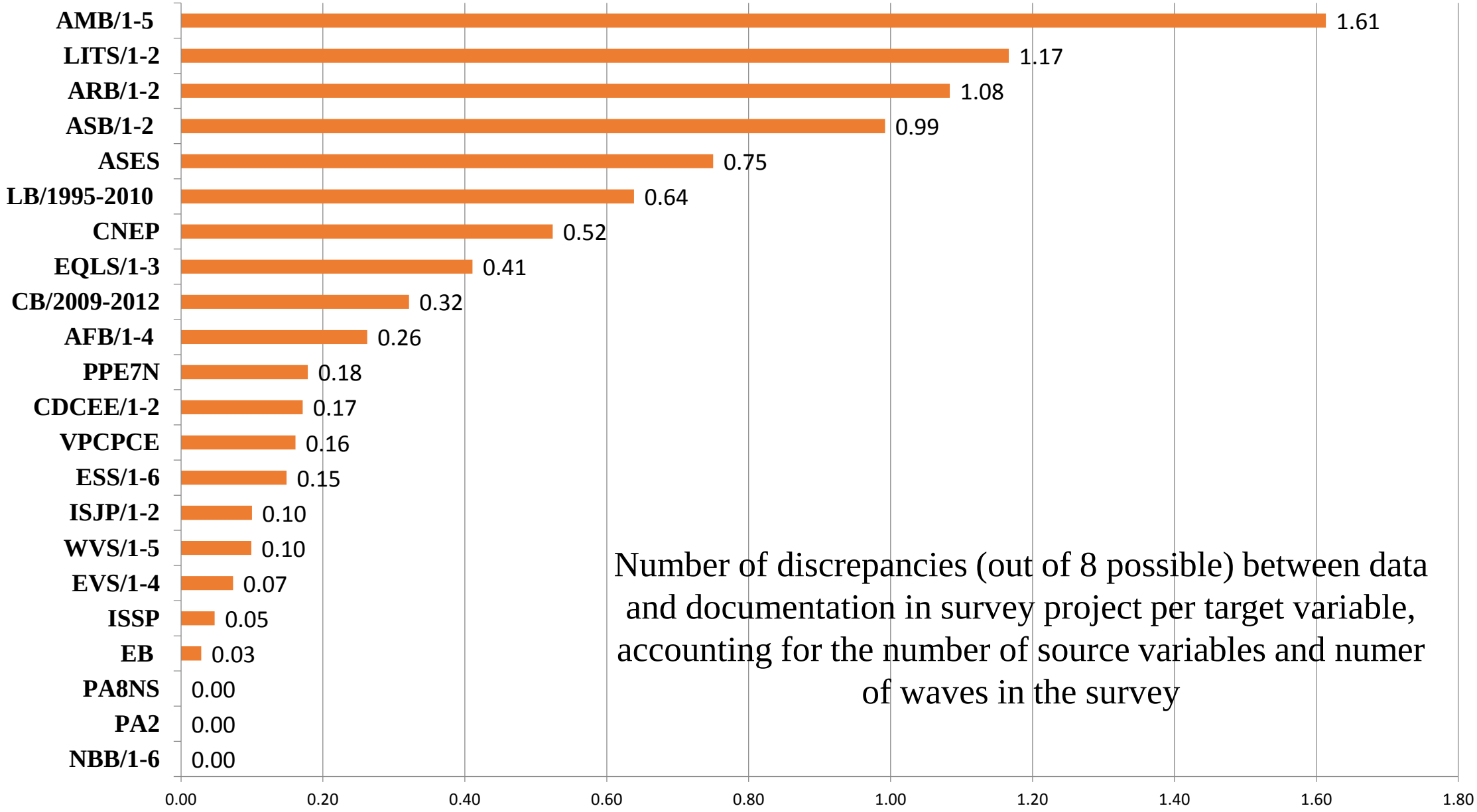
# Discrepancies per Target Variables [total =250 discrepancies]



# Types of Discrepancies [total =250 discrepancies]



# 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 |

# 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?