GPP Cleaning and Validation Protocol

EU Subnational

This is a first discussion document for internal use only.

1. Introduction

The EU Subnational project will collect public opinion surveys in 27 countries with representative samples valid to their NUTS1 or NUTS2 level (see Appendix A.1). The survey will collect people's perception about multiple topics including but not limited to Authoritarianism, Fundamental Freedoms, Corruption, Security, among others.

The quality and reliability of collected data can significantly impact the accuracy and effectiveness of the information delivered. Without a robust data validation, cleaning, and harmonization protocol in place, the risk of making erroneous conclusions, compromising operational efficiency, and undermining the trustworthiness of the outcomes is higher.

This protocol outlines specific guidelines for the cleaning, validation, and harmonization of data obtained from polling companies as part of the EU Sub-national project with the objective of ensuring the consistency, accuracy, completeness, and reliability of the projects data assets. It is important to note that these guidelines pertain solely to the data within the confines of the EU Sub-national project.

This document groups the different guidelines into four broader categories:

i. Workflow set-up
ii. Data structure validation
iii. Cleaning and harmonization
iv. Quality checks
v. TPS validation

2. Workflow set-up

2.1. SharePoint

As part of our data validation, cleaning, and harmonization protocol, all files generated and used throughout the process will be securely stored in the organization's SharePoint platform. SharePoint offers a robust and centralized document management system that ensures the accessibility, availability, and traceability of data files. This centralized storage solution not only enhances data security and confidentiality but also facilitates efficient collaboration among team members and knowledge sharing. By utilizing SharePoint as our designated storage platform, we can maintain data integrity, foster collaboration, and establish a reliable foundation for our data validation, cleaning, and harmonization efforts.

The main working directory for this data project will be located inside the following path:

World Justice Project → Research → Programmatic → EU Subnational → EU-S Data → GPP

2.2. Git version control

The use of Git and GitHub will be fundamental for this project. Git is a distributed version control system that allows us to track changes, collaborate, and manage our code and project files efficiently. GitHub, on the other hand, is a web-based hosting service that provides a platform for Git repositories, offering additional features for collaboration, code review, and project management.

By implementing Git and GitHub, we can ensure the integrity and traceability of our project files and maintain a complete history of all changes made throughout the data validation, cleaning, and harmonization process. Git's branching and merging capabilities enable multiple team members to work on different aspects of the project simultaneously, while also facilitating the integration of their contributions seamlessly. Moreover, GitHub's collaboration features, such as pull requests and code reviews, enable effective communication and feedback among team members, enhancing the overall quality and reliability of our work.

‼ Workflow suggestion: Install GitHub Desktop on your local machine.

2.3. Collection stages

i. Dummy data test (DDT)

The DDT refers to the tests and routines applied to a first batch of data received from the polling companies. This first batch consists of dummy or fake data simulated by the polling company, or in some cases, by the WJP team. No real submissions are generated during this stage and this batch serves to check the overall structure of the data that we will be receiving in the following stages and if it matches the encodings outlined in the data map file. In the case that the GPP team considers that a polling company has previously showed satisfactory scripting programming, an evaluation of the DDT will not be carried out before starting the pretest stage. However, a review of a dummy dataset helps in

ii. *Pretest Review (PTR)*

The pretest refers to a preliminary testing phase conducted before the actual survey administration. Usually refers to a sample of around 50 observations who represent the target population. The purpose of the pretest is to assess the clarity, relevance, and effectiveness of the survey questions, instructions, and response options. It helps identify any ambiguities, potential biases, or issues that may arise during data collection. During the pretest stage, we can test the first 7 steps of the protocol (see Table 1).

iii. *Midfield submission (MFS)*

If quality concerns were detected or if issues in the data structure are still present during the pretest stage, a midfield submission (MFS) is going to be asked to the polling company. This MFS will consist of a data batch of around 20% of the planned sample for the Full Fieldwork (FFW). It is worth noticing that this is a subset of the actual FFW data. This stage will allow us to do early tests on a set of data where issues were still not cleared at the end of the pretest.

iv. *Full fieldwork (FFW)*

The FFW is the phase where the survey instrument is distributed to the target population to collect their opinions and perceptions about the different factors of the Rule of Law. During this stage, careful attention is given to following the survey protocol, ensuring consistency in data collection procedures, and maintaining ethical considerations. During the pretest stage, we can test all 8 steps of the protocol (see Table 1).

Table 1: Cleaning and harmonization steps across data collection stages.

Activity	DDT	PPT	MFW	FFW
1. File conversion and UTF-8 encoding	✓	✓	✓	✓
2. Structure check	✓	✓	✓	✓
3. Country specific cleaning		✓	✓	✓
4. Logic and randomization checks		✓	✓	✓
5. Routing checks	✓	✓	✓	✓
6. Labelling		✓		✓
7. Quality checks		✓	✓	✓
8. Normalization		✓		✓

2.4. *Software Requirements*

The successful execution of the guidelines presented in this document relies on utilizing the appropriate software tools. For this project, we will primarily use the Stata Package for running the data wrangling and harmonization procedures. However, R and Git will also be applied in the background. The combination of these tools provides us with a comprehensive toolkit for implementing our data validation, cleaning, and harmonization protocol.

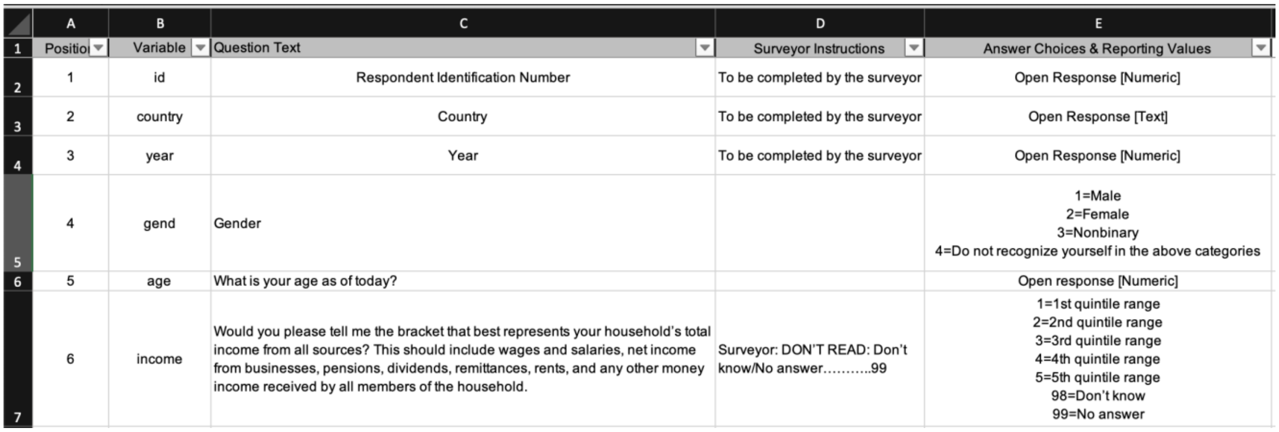
2.5. *Data documentation and support material*

To facilitate the implementation of the guidelines outlined in this document, we have prepared *(or we will prepare)* a range of support materials which main objective is to open the door for a systematic approach to data management and handling.

- Questionnaires: The survey questionnaire(s) will be the primary input for the process.

- Data Maps: Data maps provide a visual representation of the data structure, description, and encoding for the expected data that will come from the polling companies. The data map is limited exclusively to the variables contained in the official questionnaire and, therefore, *does not* include information on any variable created during the cleaning and harmonization process (e.g. country codes or NUTS code). It is important to highlight that the variable names detailed in the data map *do not necessarily* are the names that these variables will have in the final dataset (see the subsection on variable naming system below).

Figure 1: Data Map overview.

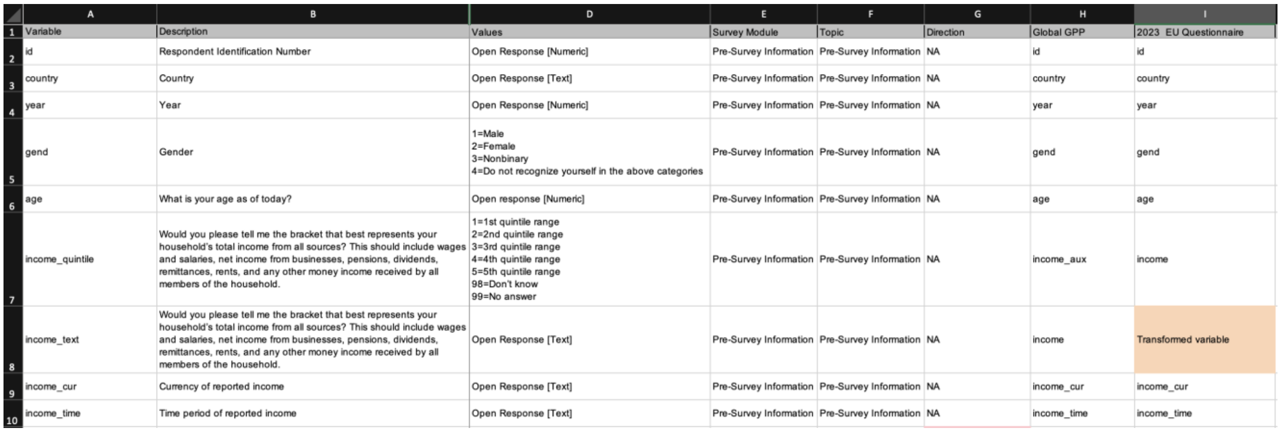


	A	B	C	D	E
1	Positio	Variable	Question Text	Surveyor Instructions	Answer Choices & Reporting Values
2	1	id	Respondent Identification Number	To be completed by the surveyor	Open Response [Numeric]
3	2	country	Country	To be completed by the surveyor	Open Response [Text]
4	3	year	Year	To be completed by the surveyor	Open Response [Numeric]
5	4	gend	Gender		1=Male 2=Female 3=Nonbinary 4=Do not recognize yourself in the above categories
6	5	age	What is your age as of today?		Open response [Numeric]
7	6	income	Would you please tell me the bracket that best represents your household's total income from all sources? This should include wages and salaries, net income from businesses, pensions, dividends, remittances, rents, and any other money income received by all members of the household.	Surveyor: DON'T READ: Don't know/No answer...........99	1=1st quintile range 2=2nd quintile range 3=3rd quintile range 4=4th quintile range 5=5th quintile range 98=Don't know 99=No answer

- Codebook: The official data set codebook serves as a comprehensive reference document that provides detailed descriptions of variables, their definitions, related topics, data types, and encoded values. It acts as a guide for data interpretation and aids in standardizing variable names, formats, and coding conventions. It is important to highlight that the codebook maps every variable included in the final dataset, *including those that are not present in the questionnaire*. Additionally, the codebook maps the equivalent names of the dataset variables to their equivalent counterpart in the

EU-GPP questionnaire and/or Global GPP dataset. An interactive tool will be available for this project to facilitate navigation along the codebook.

Figure 2: Codebook overview.



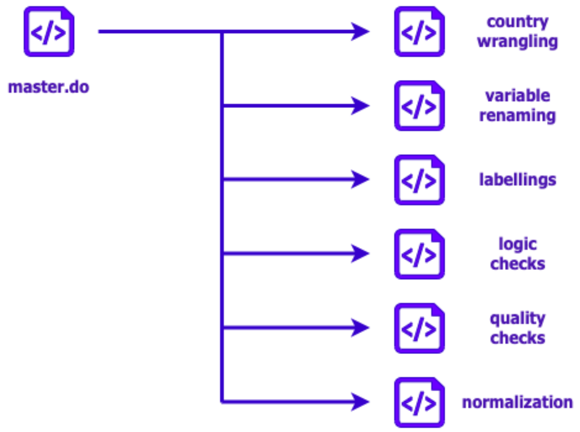
	A	B	D	E	F	G	H	I
1	Variable	Description	Values	Survey Module	Topic	Direction	Global GPP	2023 EU Questionnaire
2	id	Respondent Identification Number	Open Response [Numeric]	Pre-Survey Information	Pre-Survey Information	NA	id	id
3	country	Country	Open Response [Text]	Pre-Survey Information	Pre-Survey Information	NA	country	country
4	year	Year	Open Response [Numeric]	Pre-Survey Information	Pre-Survey Information	NA	year	year
5	gend	Gender	1=Male 2=Female 3=Nonbinary 4=Do not recognize yourself in the above categories	Pre-Survey Information	Pre-Survey Information	NA	gend	gend
6	age	What is your age as of today?	Open response [Numeric]	Pre-Survey Information	Pre-Survey Information	NA	age	age
7	income_quintile	Would you please tell me the bracket that best represents your household's total income from all sources? This should include wages and salaries, net income from businesses, pensions, dividends, remittances, rents, and any other money income received by all members of the household.	1=1st quintile range 2=2nd quintile range 3=3rd quintile range 4=4th quintile range 5=5th quintile range 98=Don't know 99=No answer	Pre-Survey Information	Pre-Survey Information	NA	income_aux	income
8	income_text	Would you please tell me the bracket that best represents your household's total income from all sources? This should include wages and salaries, net income from businesses, pensions, dividends, remittances, rents, and any other money income received by all members of the household.	Open Response [Text]	Pre-Survey Information	Pre-Survey Information	NA	income	Transformed variable
9	income_cur	Currency of reported income	Open Response [Text]	Pre-Survey Information	Pre-Survey Information	NA	income_cur	income_cur
10	income_time	Time period of reported income	Open Response [Text]	Pre-Survey Information	Pre-Survey Information	NA	income_time	income_time

- Metadata: Multiple metadata documents accompany our data sets, providing additional contextual information about the data. These documents describe the data sources, collection methodologies, processing steps, and any relevant transformations applied. Metadata documents enhance data understanding, facilitate data traceability, and ensure transparency in our data management processes.

By utilizing these support materials, we aim to establish a comprehensive framework for data validation, cleaning, and harmonization. These resources enable us to maintain data consistency, traceability, and quality throughout the entire project lifecycle. They serve as essential references for current and future team members, ensuring a standardized approach and enhancing the reliability and usability of our data assets.

2.6. Modular programming

Modular programming is a structured programming technique that facilitates the writing and development of large pieces of code by decomposing the code into smaller pieces. This technique was already implemented into the GPP cleaning routines of WJP. However, our goal is to decompose the cleaning routines further by following the structure below:



The logic behind this structure is that the analyst in charge of the country data cleaning will only have to fill some parameters in the master do file such as country name and data stage. Once these parameters have been filled, the do file can work for every incoming data set without having to duplicate this master do file or the common tasks for every incoming dataset received. Nevertheless, the analyst in charge will have to create a country specific wrangling do file to harmonize the data and allow the smooth running of all the other tasks down the pipeline.

```
*--- Parameters:

*------ (a) For which country are we going to run this dofile?
global country_name           "INSERT COUNTRY NAME HERE"
global cname                  "INSERT_COUNTRY_NAME_NOSPACES_HERE"

*------ (b) What data stage is this?
global dataStage "INSERT_DATA_STAGE_HERE"

*------ (c) Year
global year INSERT_YEAR_HERE
```

Prior to run the master_{year}.do, the analyst in charge of the country cleaning has to make sure that the input data already complies with the guidelines on data conversion and string encoding below.

3. Step 1: File conversion and UTF-8 encoding

3.1. File conversion to DTA

Usually, polling companies send their data as *SAV* data files. Given that we will be performing the cleaning and harmonization mainly through the *Stata Statistical Package*® platform, we need to convert the data that we will be receiving from the polling companies to *DTA* format. To do this, we can use the import features from Stata as follows:

```
version 17.1
cd “path_to_where_original_data”
import spss using "dataset_name.sav”
save "path_to_saving_directory/{country}_{year}_raw.dta"
```

As a general rule, the original data sent by the polling company can be found in the EU Subnational GPP directory within the EU Subnational project SharePoint. However, once the data has been transformed, the new version (in DTA format) should be saved in the EU-S Data directory with the following name system: `{country_name_nospaces}_{2023}_raw.dta`.

3.2. *String Encoding*

Given that the GPP questionnaires contain open-ended questions, and we are also expecting the polling companies to fill in some variables using text strings, it is important to include some guidelines on the harmonization of text data.

Working with text data is different than working with numeric data because a single letter can be encoded in multiple different ways depending on the string encoding system used in which the data was saved. Additionally, within our sample of countries, we find three countries with alphabets different to the Latin alphabet (Cyprus, Bulgaria, Greece). In the same manner, multiple countries reflect common use of special characters that extends the Latin alphabet. Therefore, it is essential to ensure uniformity and compatibility of the individual datasets received by each polling company to avoid issues on the final merged dataset.

To achieve this, it is recommended to convert all text data to UTF-8 encoding. UTF-8 is a widely adopted character encoding that supports a broad range of characters from different languages and scripts. By standardizing on UTF-8 encoding, we can effectively handle and process text data containing characters from diverse languages, ensuring compatibility, and avoiding potential data corruption or misinterpretation caused by incompatible encoding systems.

To convert a DTA file to UTF-8 using Stata, follow these steps:

```
version 15.1
clear
cd “path_to_where_original_data”
unicode analyze "{country}_raw.dta"”
unicode encoding set “encoding_set_name”
unicode translate “{country}_raw.dta", transutf8
```

It is worth noticing that you need to know in which encoding system is the dataset by default. The command `unicode analyze` will try to guess the encoding system. However, if this doesn’t work, we strongly suggest using the GPP Copilot App to discover the string encoding.

4. Step 2: Data structure checks

The main objective of the structure checks is to evaluate how the data is structured according to the data map shared to the polling company. For this protocol, the checks try to answer the following questions:

- Are all the variables listed in the data map present in the data set?
- Are these variables correctly named?
- Do these variables have the correct format (numeric, continuous, text)?
- Are the answers correctly encoded according to the data map?
- Are there any encoded answer that surpasses the expected value range?
- Is there any variable present in the data but not in the data map?

The analyst is welcome to use the GPP Copilot App to run the data structure checks. The app will run a quick analysis trying to answer these 6 questions for every data set submitted. Nevertheless, the App will not fix any potential issue found. Therefore, the analyst will have to include these potential issues in the country specific cleaning/wrangling routine.

5. Step 3: Country specific cleaning

The primary goal of the country-specific cleaning routine is to achieve data harmonization by transforming the data received from various polling companies into a "*clean*" dataset that can be seamlessly merged with data from other countries. This process aims to eliminate any potential conflicts arising from data inconsistencies, logical discrepancies, or encoding variations.

Since the data from each of the 27 countries included in the merged dataset may originate from different polling companies, this part of the process is highly tailored to each dataset received. Customized cleaning measures are applied to address specific issues unique to each dataset, ensuring that the data is standardized, coherent, and compatible for successful merging. By meticulously cleaning and harmonizing the country-specific data, we can establish a robust and unified dataset ready for comprehensive cross-country analysis without compromising data integrity or creating conflicts.

5.1. Dropping additional variables

If, during the data structure checks, you find variables added by the polling company that are not listed in the data map, we proceed to drop these variables from the data. Possible examples of added variables include country, *SSU*, id. These are informative variables added by the polling company for reference and we can drop them.

```
*--- Dropping variables added by the company:
drop country year id
drop {variable_list}
```

5.2. Adding general information

The following variables are added on our side from the beginning:

- country_name_off: Official country name according to the 2022 edition of the *Statistical regions in the European Union and partner countries* document. It is important to note that the official country names are not necessarily expressed in characters from the Latin alphabet.
- country_name_ltn: Unofficial country name in the Latin alphabet as they are listed in the ISO Browsing Platform (OPB). Visit the platform in this link. These values will be extracted from the master dofile.
- country_code_nuts: Official country code according to the 2022 edition of the *Statistical regions in the European Union and partner countries* document. It is important to note that this country code is a two-letters code different to the three-letter ISO code used by the index team in the Global GPP data. The codes are also available in Appendix A.1.
- country_code_iso*:* Standard three-letters code (Alpha-3) country code for each country. The ISO-3166 codes can be found in this link.
- nuts_ltn: Official name of the subnational NUTS region according to the 2022 edition of the *Statistical regions in the European Union and partner countries* document. Name is registered using the Latin alphabet.
- nuts_id: ID code of the subnational NUTS region according to the 2022 edition of the *Statistical regions in the European Union and partner countries* document.
- year: Year of data collection.
- method: Data collection methodology, it can be either "*Face-to-Face*", "*Online*" or "*Telephone*".
- id: Unique identifier per country dataset. It is equal to the row number of the observation in the data.
- country_year_id: Unique identifier for the merged dataset. It is equal to the concatenation of the following variables: country_name_ltn, year, and id.
- income_group: Income group classification according to the World Bank. Possible income groups are High (H), Upper-Middle (UM), Lower-Middle (LM) or Low (L). Classifications for the 2023 fiscal year are available in this link.

All this information will be stored as an excel spreadsheet and as a Stata DTA file for a quick merge using the global value of the country name defined at the beginning of the master dofile.

```
*--- Generating new id
g id = _n

*--- Adding country information
g country_name_ltn = "${country}"
merge m:1 ${country} using "general_info.dta", nogen

*--- Unique identifier for merged dataset
egen country_year_id = concat (country_name_ltn year id), punct("_")
```

‼ Note: Data under the EU Subnational Project will be harmonized in accordance with the project needs. A second dataset with the appropriate information to be merged with the Global GPP data will be produced afterwards. See Section 12 below.

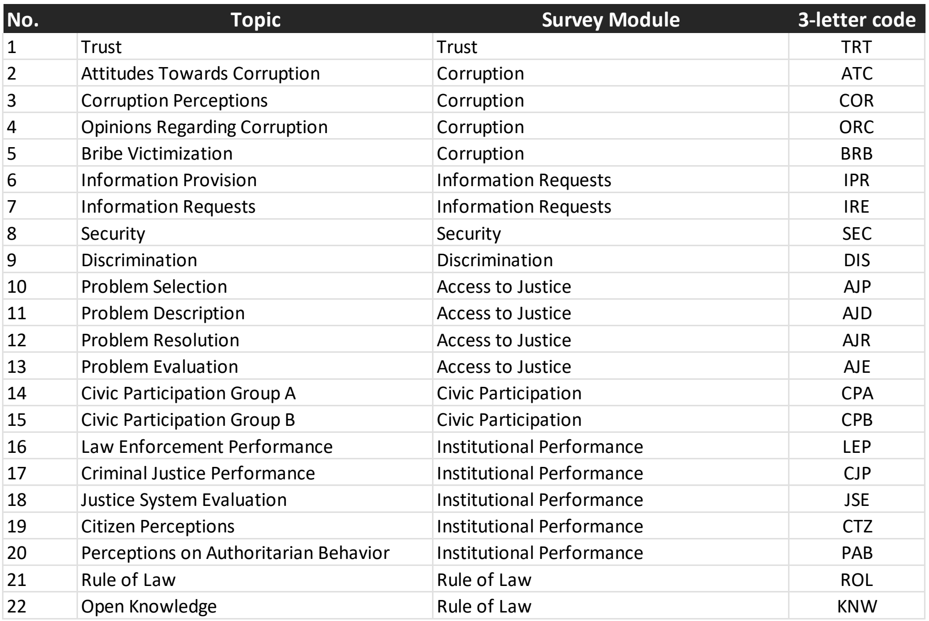
5.3. Renaming and recoding variables

A significant portion of the country-specific cleaning process involves renaming and recoding variables to ensure their compatibility for merging purposes. During this step, the analyst will have to code a battery of commands to ensure the following:

- Any flagged issue during the structure checks is fixed.
- Any potential mismatch between the data map and the Codebook is resolved.
- All variables are re-named according to the Codebook.
- If a shorter/modified questionnaire was applied in a specific country, the empty variables are created.

i. Variables naming system

For the European Union Subnational project, we are adopting a semantic naming system. The name will summarize the topic to which the question is associated, the question keywords and an optional extension to better clarify the target, institution, or geographical extension of the question. The survey has been divided into 9 thematic modules and 22 topics. See the table below:



No.	Topic	Survey Module	3-letter code
1	Trust	Trust	TRT
2	Attitudes Towards Corruption	Corruption	ATC
3	Corruption Perceptions	Corruption	COR
4	Opinions Regarding Corruption	Corruption	ORC
5	Bribe Victimization	Corruption	BRB
6	Information Provision	Information Requests	IPR
7	Information Requests	Information Requests	IRE
8	Security	Security	SEC
9	Discrimination	Discrimination	DIS
10	Problem Selection	Access to Justice	AJP
11	Problem Description	Access to Justice	AJD
12	Problem Resolution	Access to Justice	AJR
13	Problem Evaluation	Access to Justice	AJE
14	Civic Participation Group A	Civic Participation	CPA
15	Civic Participation Group B	Civic Participation	CPB
16	Law Enforcement Performance	Institutional Performance	LEP
17	Criminal Justice Performance	Institutional Performance	CJP
18	Justice System Evaluation	Institutional Performance	JSE
19	Citizen Perceptions	Institutional Performance	CTZ
20	Perceptions on Authoritarian Behavior	Institutional Performance	PAB
21	Rule of Law	Rule of Law	ROL
22	Open Knowledge	Rule of Law	KNW

For example, question q2d reads as follows:

"What is your opinion about the following behaviors? Is it always acceptable, usually acceptable, sometimes acceptable, or not acceptable? An elected official taking public funds for private use."

Given that question q2d is part of the Corruption Module and it is listed to the Attitudes towards Corruption set, then its name is formed as follows:

ATC_embezz_priv

Where **ATC** is the three-letter code for the Attitudes towards Corruption set, **embezz** is the keyword used to describe the question objective and **priv** is the optional extension used to complement the target of the question and make it differentiable for other questions about embezzlement.

‼ Note: Data under the EU Subnational Project will be harmonized in accordance with the project needs. A second dataset with the corresponding naming system of the Global GPP data will be produced afterwards. See Section 12 below.

ii. Special cases

Some variables need to be either further modified, follow specific encodings, or require further tests. Below, you can find a list enumerating these special cases with a brief explanation.

- *Ethnicity* (ethn):

The values for the ethnicity variable should be encoded following the European Standard Classification of Cultural and Ethnic Groups. Display Cards for 19 European countries are also available in the methodological appendix of the European Social Survey. The resulting variable will be numeric with assigned value labels. A new variable called "ethn_class" will be created for each country. This new variable will classify the stated ethnicities with the following value labels from 1 to 3:

1. *Main ethnicity* if the stated ethnicity is part of the major ethnicity in the specified country.
2. *EU minor ethnicity* if the stated ethnicity is a minor ethnicity but also part of the European Union.
3. *Non-EU minor ethnicity* if the stated ethnicity is a minor ethnicity from outside the European Union.

- *Political party encoding* (voteintention)*:*

The values for the party affiliation variable should be encoded following the Party Coding Units established by the V-Dem project. The resulting variable will be numeric with assigned value labels. A new variable called "incpp" will be created for each country. This new variable will have a binary encoding under which, its value will be equal to 1 if the party is the incumbent political party and equal to zero otherwise. The information of the incumbent political party in each country will be available in the general_information.dta file.

- *Age* (age):

Please ensure that age is a continuous variable, no stated age is under 18 years old, and no age ranges were coded for this question.

- *NUTS Region* (nuts_name & nuts_code):

Similarly the country codes, the NUTS region name and the NUTS region should follow the conventions of the *Statistical regions in the European Union and partner countries* document. The region names and codes are also available in Appendix A.1.

- *Missing values*

Some polling companies have specific values to encode missing or empty responses such as "*.z*", "*.x*" or "*.n*". Assess this at the beginning of every country routine received and modify accordingly from Stata.

5.4. *Filling empty variables*

It is important to consider that not necessarily the same exact questionnaire was applied in all countries. The number of variables in the data might differ depending on three factors. First, there are two possible lengths for the questionnaires: the full and abridged version. Depending on the proposals made by the polling companies and budget considerations, one out of these two versions will be applied.

Second, countries in which the data was collected using a Computer Assisted Personal Interview approach (CAPI) will usually have more variables than countries where a Computer Assisted Web Interview was applied due to the nature of some questions that can only be applied in face-to-face settings.

Third, depending on the socio-economic and political context of each country, polling companies might request to drop specific variables from the questionnaire to ensure a smooth overall collection.

To facilitate the harmonization and validation process, it is highly important to ensure that all variables recorded in the codebook are present in the data. Therefore, it is important to record which variables were not collected and generate empty columns – missing values – for them.

5.5. *Ordering the data*

One last step before saving the country routine is to ensure that the variables in the data are ordered according to the codebook for easy navigation across the dataset.

5.6. *Saving the country routine*

It is important that all the changes applied during this step are grouped and follow the order of the modules in the questionnaire. Please use this dofile template to write your Stata routine. Once that all the changes have been recorded in the dofile, the country specific routine should be saved as {country}_wrangling{year}.do in a separate dofile in the following path:

→ EU Subnational → EU-S Data → GPP → 2. Code → Country-Wrangling → $dataStage

6. Step 4: Logic and routing checks

Logic and routing checks are essential procedures that need to be implemented after harmonizing the data received from the polling company. These checks are standard across datasets and aim to ensure the integrity and consistency of the collected responses. The logic tests assess whether the answers align with a coherent and non-contradictory narrative. There are six logic checks in place:

a. Age restriction is in place (age): All respondents should be within the age range of 18 and 99 years old.
b. Automatic selection of the legal problem (q17): The selected problem in the Access to Justice module should only consider problems which had a recorded severity of at least 4 out of 10. Problems with registered severities below 4 are only eligible for the automatic selection if and only if not a single problem had a severity of at least 4.
c. We flag observations reporting having had to skip their job for over a year (q38e).
d. We flag observations reporting more than 100 medical visits (q38g_1).
e. We flag observations reporting to have been hospitalized for over a year (q38h_1).
f. We flag observations reporting a household size of 16 people or more (A1).

A parallel set of routing tests try to assess if the answers of the respondents followed the skips displayed in the questionnaire. In this regard, we assess two routes:

- First, we test the skip route, meaning that respondents that should have skipped a specific question, should have missing values as response to that question.
- Second, we test the non-skip route, meaning that respondents that should have answer a question after a skip, are indeed answering that question.

By conducting these checks, we can identify any inconsistencies or illogical patterns in the data, allowing us to address and rectify potential issues, thereby enhancing the overall quality and reliability of our dataset. A dofile with all these tests will be available for implementation in the Codes folder with the name: logic_checks{year}.do.

7. Step 5: Randomization checks

The GPP questionnaire consists of multiple modules that are divided into two groups: Group A and Group B. This division is implemented to reduce the length of the questionnaires while retaining all variables. The assignment of respondents to these groups is intended to be random, ensuring that each respondent has a 50% probability of being assigned to either module whenever a split module is encountered. Randomization checks are performed to evaluate the effectiveness of this randomization procedure within the data. In the current edition of the GPP, we are assessing two specific conditions:

- First, that the probability of being assigned to Group A or Group B is equal for all respondents. For this, we compare the frequency tables of the first question in each group.
- Second, that the probability of being assigned to Group A or Group B is not conditional on the previous assignments. For example, if a given respondent was assigned to group A in the institutional performance module, its probability of being assigned to the same group in the rule of law module remains at 50%. For this, we compare the cross-frequency tables between different groups in different modules.

A dofile with all these tests will be available for implementation in the Codes folder with the name: logic_checks{year}.do, same file were the logic and routing checks are placed.

8. Step 6: Labelling

Variable labels play a crucial role in providing a concise description of the question being asked. Therefore, to ensure harmonized descriptions, we will be dropping labels written by the polling company and replace them by a set of standardize labels. This applies to both, variable descriptions, and value labels.

```
*--- Drop ALL labels added by the company:
foreach x of varlist * {
  label var `x' ""
}
label drop _all

*--- Applying standard labels:
do "${path2dos}/var_labels${year}.do"
do "${path2dos}/val_labels${year}.do"
```

To ensure compatibility with Stata's maximum character limit for variable labels (80 characters), we will adhere to the guideline of keeping the variable label within this limit. However, it is important to note that the complete text of each question will be available as a question-specific note in the dataset file. Both, the full phrasing, and the shortened 80-characters long description will be available in the data codebook.

9. Step 7: Quality checks

Implementing in-survey quality checks is an essential approach to ensure the integrity and reliability of survey responses. These checks serve as indicators of whether respondents are thoughtfully engaging with the survey. Quality checks can take various forms, ranging from straightforward filters that exclude respondents who complete the survey too quickly to more sophisticated methods, such as requesting participants to validate their earlier responses. Recognizing that the quality of a survey is dependent on the data it generates, incorporating quality checks is considered a best practice in survey validation.

9.1. Measuring missing values

As part of our data validation, cleaning, and harmonization protocol, we will systematically measure the presence of missing values within each observation in the data. Missing values can significantly impact the integrity and reliability of our analyses. To address this, we will calculate the number of missing values per observation and establish a predetermined threshold. Observations exceeding this threshold will be flagged for further investigation and potential corrective actions. By implementing this approach, we can proactively identify and address data incompleteness issues, ensuring the accuracy and completeness of our dataset and ultimately enhancing the quality and reliability of our analyses.

9.2. Answering flags

i. Speeder flag

Speeder Flag is used to identify respondents who have completed the survey at an unusually fast pace. This check aims to identify participants who may not have carefully considered the questions or provided thoughtful responses. To perform this check, we will make use of the survey length variable and its derived versions to compare against pre-determined thresholds. Observations that were completed significantly below these thresholds will be flagged.

ii. Straight-lining flag

Straight-lining Flag is used to detect respondents who consistently select the same response option (e.g., always choosing the first option) without considering the content of the question. This indicates a lack of attention or engagement with the survey. To perform this check, we will estimate the number of straight answers supplied by a respondent. The observations with unusual long streaks will be flag.

iii. Conflicting answers

Conflicting answers occur when a respondent provides contradictory responses within the survey. These inconsistencies may be within the same question or across related questions. A few questions have been placed in the questionnaire with this objective in mind. We will estimate the overall percentage in the dataset that has conflicting answers.

9.3. Difficulty index

For Face-to-Face surveys, the questionnaire leaves a serious of questions to be answered by the surveyor. These questions are aimed to assess how difficult and, as a result, how reliable are the answers given by the respondents. We will use these variables to calculate a "*Difficulty Index*" and assess the overall dataset comparing the results with other countries.

In Face-to-Face surveys, the questionnaire includes a series of questions specifically designed to be answered by the surveyor. These questions serve the purpose of evaluating the level of difficulty associated with obtaining responses from respondents, which, in turn, provides insights into the reliability of the answers provided. By using these variables as inputs, we can calculate a "*Difficulty Index*" through a Principal Component Analysis. Comparing the results with other countries allows for a comprehensive evaluation of data quality and consistency across different contexts.

10. Step 8: Saving the data

Save the dataset as a Stata data file (DTA) with the name `${country}_clean.dta` in the designated path file.

```
save "/.../${country}_clean.dta", replace
```

11. Normalization and Aggregation

Unlike the Global GPP framework, in our data analysis approach, we will not be aggregating survey data into theoretical scores per pillar or sub-pillar. Instead, we will focus on using normalized and aggregated data for conducting validation checks using third-party sources. While we understand the potential benefits of creating theoretical scores, our primary objective is to ensure the accuracy and reliability of our analysis by comparing our data with external benchmarks.

By normalizing and aggregating the data, we can effectively perform these validation checks and evaluate the alignment and consistency of our findings with trusted external sources. This approach allows us to maintain transparency and ensure the integrity of our analysis while leveraging the benefits of normalized and aggregated data in the validation process. However, these normalized data points will not be saved in the merged dataset to maintain its complexity to a minimum.

‼ Note: Given that some variables from the EU-S GPP data will be used to calculate the global scores of the Rule of Law Index, a second dataset with the normalized scores will be available for merging with the Global GPP data. See Section 12 below.

11.1. Normalization

Normalization refers to the process of transforming individual data points within a single variable to a common scale or distribution that spreads across a set of indicators. This is typically done to facilitate meaningful comparisons between variables that may have different units or ranges. There are multiple techniques for normalizing variables. For the purposes of this project, we will be using the min-max scaling approach to normalize indicators within a [0-1] scale. For this, we will be applying the following formula:

$$\tilde{x} = \frac{x - x_{min}}{x_{max} - x_{min}},$$

Where $\tilde{x}$ is the normalized data point, x is the data point in its original scale, x_{min} is the minimum possible value for the variable in its original scale and x_{max} is the maximum possible value for the variable in its original scale.

Because variables can either be positively or negatively related to the Rule of Law, it is extremely important to adjust the direction of all the variables to a common ground. For this, we will be multiplying the data points by either 1 (if the variable is positively correlated) or by -1 (if the variable is negatively correlated) prior to applying the min-max scaling technique.

11.2. Aggregation

During the quantitative checks for the validation of the data, it might be necessary to aggregate normalized data. When necessary, a simple weighted arithmetic mean procedure will be applied.

12. Harmonization with the Global GPP Data

Once that a ready-to-be-merged dataset has been saved. A special Stata dofile can be run over this file to create a ready-to-be-merged file for the Global GPP dataset. This dofile will have the following tasks:

- Drop variables proper from the EU-S project.

- Add required variables for the Global routines.
- Rename variables to ensure harmonization.
- Recode special variables to ensure harmonization with global GPP data.
- Recode the DK/NA values.
- Add normalized scores.

13. Steps Checklist

Before wrapping up the cleaning and validation process for a given country data, please answer the following questions:

- □ Has the data been harmonized into UTF-8 encoding?
- □ Are all the variables from the codebook present in the data?
- □ Is there any variable in the data that is not present in the codebook?
- □ Do all the variables have encoded answers within their expected range?
- □ Do all the special case variables follow their specific guidelines?
- □ Have the logic and routing checks been performed?
- □ Has any potential issue flagged by the logic and routing checks been fixed?
- □ Have the randomization checks been performed?
- □ Have the quality checks been performed?
- □ Do all the variables have proper 80-characters long labels?
- □ Is the saved dofile up and running without issues?

15. Appendix

A.1 List of countries and NUTS regions

Country	NUTS ID	Name Latin
Austria	AT	Austria
	AT1	Ostösterreich
	AT2	Südösterreich
	AT3	Westösterreich
Belgium	BE	Belgium
	BE1	Région de Bruxelles-Capitale/Brussels Hoofdstedelijk Gewest
	BE2	Vlaams Gewest
	BE3	Région wallonne
	BG3	Severna i Yugoiztochna Bulgaria
	BG4	Yugozapadna i Yuzhna tsentralna Bulgaria
Bulgaria	BU	Bulgaria
	BU0	Bulgaria
Cyprus	CY	Cyprus
	CY0	Kýpros
Czechia	CZ	Czechia
	CZ01	Praha
	CZ02	Střední Čechy
	CZ03	Jihozápad
	CZ04	Severozápad
	CZ05	Severovýchod
	CZ06	Jihovýchod
	CZ07	Střední Morava
	CZ08	Moravskoslezsko
Germany	DE	Germany
	DE1	Baden-Württemberg
	DE2	Bayern
	DE3	Berlin
	DE4	Brandenburg
	DE5	Bremen
	DE6	Hamburg
	DE7	Hessen
	DE8	Mecklenburg-Vorpommern
	DE9	Niedersachsen

	DEA	Nordrhein-Westfalen
	DEB	Rheinland-Pfalz
	DEC	Saarland
	DED	Sachsen
	DEE	Sachsen-Anhalt
	DEF	Schleswig-Holstein
	DEG	Thüringen
Denmark	DK	Denmark
	DK01	Hovedstaden
	DK02	Sjælland
	DK03	Syddanmark
	DK04	Midtjylland
	DK05	Nordjylland
Estonia	EE	Estonia
	EE0	Eesti
Greece	EL	Greece
	EL3	Attiki
	EL4	Nisia Aigaiou, Kriti
	EL5	Voreia Elláda
	EL6	Kentriki Elláda
Spain	ES	Spain
	ES1	Noroeste
	ES2	Noreste
	ES3	Comunidad de Madrid
	ES4	Centro (ES)
	ES5	Este
	ES6	Sur
	ES7	Canarias
Finland	FI	Finland
	FI19	Länsi-Suomi
	FI1B	Helsinki-Uusimaa
	FI1C	Etelä-Suomi
	FI1D	Pohjois- ja Itä-Suomi
	FI20	Åland
France	FR	France
	FR1	Ile-de-France
	FRB	Centre — Val de Loire
	FRC	Bourgogne-Franche-Comté

	FRD	Normandie
	FRE	Hauts-de-France
	FRF	Grand Est
	FRG	Pays de la Loire
	FRH	Bretagne
	FRI	Nouvelle-Aquitaine
	FRJ	Occitanie
	FRK	Auvergne-Rhône-Alpes
	FRL	Provence-Alpes-Côte d’Azur
	FRM	Corse
Croatia	HR	Croatia
	HR02	Panonska Hrvatska
	HR03	Jadranska Hrvatska
	HR05	Grad Zagreb
	HR06	Sjeverna Hrvatska
Hungary	HU	Hungary
	HU1	Közép-Magyarország
	HU2	Dunántúl
	HU3	Alföld és Észak
Ireland	IE	Ireland
	IE04	Northern and Western
	IE05	Southern
	IE06	Eastern and Midland
Italy	IT	Italy
	ITC	Nord-Ovest
	ITF	Sud
	ITG	Isole
	ITH	Nord-Est
	ITI	Centro (IT)
Lithuania	LT	Lithuania
	LT01	Sostinės regionas
	LT02	Vidurio ir vakarų Lietuvos regionas
Luxembourg	LU	Luxembourg
	LU00	Luxembourg
Latvia	LV	Latvia
	LV00	Latvija
Malta	MT	Malta
	MT00	Malta

Netherlands	NL	Netherlands
	NL1	Noord-Nederland
	NL2	Oost-Nederland
	NL3	West-Nederland
	NL4	Zuid-Nederland
Poland	PL	Poland
	PL2	Makroregion południowy
	PL4	Makroregion północno-zachodni
	PL5	Makroregion południowo-zachodni
	PL6	Makroregion północny
	PL7	Makroregion centralny
	PL8	Makroregion wschodni
	PL9	Makroregion województwo mazowieckie
Portugal	PT	Portugal
	PT1	Continente
	PT2	Região Autónoma dos Açores
	PT3	Região Autónoma da Madeira
Romania	RO	Romania
	RO1	Macroregiunea Unu
	RO2	Macroregiunea Doi
	RO3	Macroregiunea Trei
	RO4	Macroregiunea Patru
Sweden	SE	Sweden
	SE1	Östra Sverige
	SE2	Södra Sverige
	SE3	Norra Sverige
Slovenia	SI	Slovenia
	SI03	Vzhodna Slovenija
	SI04	Zahodna Slovenija
Slovakia	SK	Slovakia
	SK01	Bratislavský kraj
	SK02	Západné Slovensko
	SK03	Stredné Slovensko
	SK04	Východné Slovensko