The Triangle of polarization, political trust and political communication: Understanding its dynamics in contemporary democracies.

(TRI-POL) (2019-2022)

ITALY

Data protocol FOR THE PASSIVE METER TRACKER

Funding

This research was funded by two competitive grants. First, the Spanish Ministry of Economy and Competitivity. Ministerio de Economía y Competitividad, Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia PID2019-106867RB-I00 /AEI/10.13039/501100011033 (2020-2024), Principal Investigador: Mariano Torcal). Second, by the Fundación BBVA, Ayudas a Equipos de Investigación Científica en Economía y Sociedad Digital 2019 (2020-2022). The views expressed herein are those of the authors and are not necessarily those of these two funding agencies. The PI of the project is also grateful for the funding provided by the Advance Research Fellowship Programme ICREA, funded by the Catalonian Government.

Index

Index of Tables

TRI-POL 2021-2022 Behavioral Dataset

Technical Information

1. Citation, Research Team and Contact

Citation

This dataset is provided free of charge for all those who wish to use it. Designing this study, retrieving the data, cleaning it, and preparing it for public use meant a lot of work. We are therefore grateful for your acknowledgment of our efforts by citing the database when you use it. The suggested citation is the following:

Torcal, Mariano; Emily Carty, Oriol J. Bosch, Josep Comellas, Zoe Thomson and Danilo Serani (2022). "The Triangle of Polarization, Political Confidence and Political Communication: Understanding its Dynamics in Five Contemporary Democracies", *Data in Brief*,

Research Team

Mariano Torcal (Universitat Pompeu Fabra).
Oriol J. Bosch (London School of Economics)
Emily Carty (Universidad de Salamanca),
Melanie Revilla (Universitat Pompeu Fabra),
Ryan Carlin (Georgia State University)
Greg Love (University of Mississippi)
Noam Lupu (Vanderbilt University)
Pedro Magalhaes (University of Lisbon)
Matias Bargsted (Universidada Católica de Chile)
Carolina Segovia (Universidad Diego Portales)
Danilo Serani (Universidad de Salamanca)
Josep Maria Comellas (Universitat Pompeu Fabra),
Zoe Thomson (Universitat Pompeu Fabra)

Contact

Oriol Bosch:
Mail: O.Bosch-Jover@lse.ac.uk

Additionally, you can contact:
Mariano Torcal
Mail: mariano.torcal@upf.edu

2. Data Description

Overview

The TRI-POL dataset is a micro-level online panel survey in five countries: Argentina, Chile, Italy, Portugal and Spain among their respective voting age population comprised of three waves carried out over a six-month period between late October 2021 and May 2022 (the detailed timing of each wave will be presented in Table 1). In addition, the project comprises a series of survey experiments, embedded in the different waves, regarding social exposure, polarization framing and social sorting. This dataset and project also includes variables based on tracking respondents behaviour collected by a passive meter using a software that the interviewees installed on their mobile devices.

The following protocol contains technical information concerning the passive tracking methodological approach.

Files

15 TRI-POL Behavioural data collected with Passive Meter, one for each country and wave (Stata 17.0 files)

3. General Tracking Design

General approach

As defined in the survey data protocol, the surveys were administered by Netquest to a sample of participants from their metered panels, selected using non-probability quota sampling. Panellists from the Netquest metered panels have – knowingly and consensually - digital tracking solutions installed in at least one of their devices, which allows to complement their survey answers with information about their online behaviours.

Overall, we were able to obtain online behavioural information from 842 of the 1,231 participants who completed the first survey wave. Challenges were faced when filling some of the specific cross-quotas with participants from the metered panel. This required supplementing in some cases with non-metered panellists, hence, the 31.6% of participants without tracking information.

TRI-POL researchers did not have access to the raw data with information about all URLs and apps visited by panellists, and their respective timestamps, to minimize any potential ethical concern linked with this project. Alternatively, a list of variables and guidelines on how to compute them was developed and sent to Netquest, for them to implement. The guidelines can be checked here and here. Netquest created and delivered several anonymized structured datasets, which complied with our specifications. Those databases were then processed by members of the TRI-POL project to create the databases here described i.e., three separate databases for each country, one for each wave.

Tracking approach

Online behavioural data was collected for the 15 days prior and posterior of participants starting each survey wave. The meter captured each URL (or app for mobile devices) accessed by the panellists, with timestamps for when the panellists first visited the URL, and the number of seconds in which the URL remained active in the browser. A URL was considered active when it was the one being displayed in the browser, meaning that other URLs that may be open in other tabs were not considered to be active. The number of active seconds was measured as the time between the URL (or app) first becoming active in the browser (i.e., displayed to the respondent) and a different URL (or app) becoming active in the browser.

Participants were tracked on iOS and Android mobile devices, and Windows and MAC computers, using the tracking solutions provided by Wakoopa (https://www.wakoopa.com/). Specifically, Windows and MAC devices were tracked with desktop apps and/or web browser plug-ins, android devices through apps and iOS devices through manually configured proxies. Information about which technologies were used to track each participant was requested to Netquest, which is provided in the databases. Table 1 provides more information about the capabilities and limitations of the different technologies used.

Table 1 Data collectable by tracking technology and target, for Wakoopa

		PC app	**PC plug-ins**			**Android SDK**	**iOS proxy**
			Chrome	**Firefox**	**Safari**		
Online tracking							
URLs	Http traffic	Yes	Yes	Yes	Yes	Yes	Yes
	Https traffic	No	Yes	Yes	Yes	Yes	No
	Incognito sessions	No	Yes	Yes	Yes	Yes	No
	HTML	No	Yes	Yes	Yes	No	No
	Time stamps	Yes	Yes	Yes	Yes	Yes	Yes
Apps	App name	-	-	-	-	Yes	Yes
	App usage start time	-	-	-	-	Yes	Yes
	App usage duration	-	-	-	-	Yes	Estimated
	Offline apps	-	-	-	-	Yes	No
	In-app behaviour	-	-	-	-	No	No
Search terms	Search terms	Yes	Yes	Yes	Yes	Yes	No
Device information							
Device type	E.g. desktop	Yes	Yes	Yes	Yes	Yes	Yes
Device brand	E.g. Xiaomi		No	No	No	Yes	Yes
Device model	E.g. S9	No	No	No	No	Yes	Yes
Operating system	E.g. iOS	Yes	Yes	Yes	Yes	Yes	Yes
OS version	E.g. 10.1.2	No	No	No	No	Yes	Yes
Internet provider	E.g. Voxi	No	No	No	No	Yes	Yes

Variables Defined

As previously discussed, Netquest computed a series of predefined variables following our instructions. All these variables measured the daily number of visits or seconds spent on a set of given webpages or groups of webpages, as well as to specific content (e.g., political articles) within those webpages.

Specifically, we asked Netquest to measure the number of visits or time spent on the top 50 the most popular news media outlets in Italy (according to https://tranco-list.eu/) and social media. Within those, we measured the visits and seconds spent on URLs defined as opinion articles, news in general and national, regional, international and political news. In addition, we created variables measuring the visits and seconds spent on specific Twitter profiles (the ones used for the experimental design). The URLs defined to create all the variables in the database for Italy can be checked here.

4. Methodological Information about the Tracking Approach

The tracking approach was designed following the Total Error framework for digital traces collected with Meters (TEM, Bosch and Revilla, 2022a). The TEM provides guidelines on how to better design, analyse and report metered data. Below we provide more information on how the TEM was applied when building the database. Following the transparency best practices suggested by the TEM, we also provide empirical evidence for some data quality indicators. More in-depth explanations about our approach can be found in Bosch and Revilla (2022a; 2022b).

Tracking Undercoverage

All the variables in this database were intended to measure behaviours at the individual level (e.g., how much time someone spends reading news articles). Nonetheless, metered data measures are the combination of all the behaviours that an individual does through all the devices, web-browsers, and apps that they use, and all the networks that they connect to (Bosch and Revilla, 2022). Thus, to observe the complete behaviour of an individual, meters must be installed on all the specific *targets* that they use to go online. This might not always be possible to achieve because of issues such as technology limitations or unwillingness of participants of installing tracking technologies in all the targets that they use. In those cases, only a partial image of a participant's online behaviour is observed, which can lead to errors such as underestimation of univariate estimates. This is known as tracking undercoverage.

Given that we were using a sample of participants who were already being tracked by Netquest, it was out of our control to make sure that every participant was being fully covered. In those cases, and for transparency's sake, the TEM proposes reporting the proportion of participants being affected by tracking undercoverage, similarly with what is done with nonresponse rates.

After applying the method proposed by Bosch and Revilla (2022b), we found the following proportion of individuals with at least one device not being tracked, for waves 1 and 3.

Table 2 Proportion of participants undercovered in terms of device, in all countries for waves 1 and 3

	Italy		**Portugal**		**Spain**		**Argentina**		**Chile**	
	W1	**W3**	**W1**	**W3**	**W1**	**W3**	**W1**	**W3**	**W1**	**W3**
Device	76.1	76.7	76.5	75.3	70.3	66.9	70.0	67.9	73.7	72.7
N	842	688	818	675	992	844	1,127	848	958	693

Note: unweighted proportions

These results agree with what previous research has found (see Revilla, Ochoa and

Loewe, 2017; Pew Research Centre, 2020), indicating that our project is in line with most research done so far with metered data. Preliminary evidence simulating the impact of this on the TRI-POL databases suggests that this might result in univariate estimates being underestimated by around 7 to 15% (Bosch, 2022). Users should be mindful about this fact and consider simulating to what extent tracking undercoverage could affect their results (we recommend using Bosch, 2022 simulation approach, or similar).

Misclassified non-observations

Metered data is nonreactive, which means that it is not created by asking participants to provide information but by passively observing their behaviours online. The behavioural variables in the database, therefore, represent the time or number of visits that we observed individuals doing some specifically defined behaviours. Problematically, errors such as tracking undercoverage can prevent researchers from observing all the behaviours that participants do online (see Bosch and Revilla, 2022a; 2022b). Sometimes the observed behaviours might not correspond with individual's true behaviours.

In the past, research using metered data has assumed that when no information was observed for a specific behaviour (e.g., time spent visiting Facebook), this meant that the person had not done that behaviour. However, when we factorize errors into the equation, a lack of observed behaviour might also mean that the person did in fact do that behaviour, but it was not observed. Hence, when errors are present it is not possible to clearly discern when a lack of observed behaviours should be treated as real (e.g., 0 seconds visiting Facebook) or as a missing (i.e., a real behaviour happened, but we did not observe it) without auxiliary information.

Being mindful about this and knowing that our databases are not free of errors, we applied Bosch and Revilla's (2022b) approach to identify when a specific observed lack of behaviour in the TRI-POL database was true or induced by errors. This was done for the variables measuring the number of visits or seconds spent on Facebook, Twitter and the 10 most popular news media outlets in Italy (according to https://tranco-list.eu/): Repubblica, Libero, Corriere Gazzetta Del Sud, ANSA, Dgospia, La Stampa, ilsole24ore, Virgilia, GDS.

For all the other variables, we applied a simpler approach. When we knew that a participant was fully tracked, we considered their non-observations as real. When we knew that they were partially untracked, we considered their non-observations as dubious, treating them as neither real nor error induced.

In waves 1 and 3, the proportion of participants with error-induced non-observations for each of these domains was the following.

Table 3 Proportion of participants with error-induced non-observations, out of all the tracked participants

	Italy		**Portugal**		**Spain**		**Argentina**		**Chile**	
	W1	**W3**	**W1**	**W3**	**W1**	**W3**	**W1**	**W3**	**W1**	**W3**
Facebook	10.5	30.5	10.6	39.5	11.1	22.8	9.8	37.5	10.9	30.2
Twitter	23.0	15.0	17.7	15.8	14.7	17.7	16.1	25.2	21.1	26.7
Avg. News outlets	9.0	13.6	18.8	25.4	11.8	15.3	10.0	16.3	17.5	24.1
N	842	688	818	675	992	844	1,127	848	958	693

Note: unweighted proportions

TRI-POL users should consider that these proportions mean that there is a non-negligible risk of increasing the size of the estimate’s measurement errors if these participants are not excluded from the analyses. In the last section we propose ways in which researchers can address this issue.

5. Coding, Naming, and Labelling Protocols

Information in the databases follows a series of protocols to optimize the size of the database and to facilitate the users' access to and understanding of the information. The following subsections share the naming, labelling, and coding protocols employed in the TRI-POL database.

Coding of non-observations

Previously, we explained how misclassifying error-induced non-observations as real non-observations could be problematic. Additionally, we also presented our approach to identify them for some of the variables in this database. In this section we explain how we used that knowledge and information to code the non-observations in our databases. Specifically, for those variables for which we collected enough auxiliary information, we were able to code the observed non-observations as either real or induced by errors (such as tracking undercoverage). Nonetheless, many times it was not possible to identify whether non-observations for individuals affected by tracking undercoverage were true or induced by errors, hence, we also report when we did not have enough information to make a proper decision. In addition, sometimes we did not observe any behaviour of an individual because it was part of the subgroup of participants not tracked, or because it had abandoned the study in a later wave.

The coding of the different types of lack of observed behaviour was standardised for all the behavioural variables in the database, so that each type of lack of observed behaviours receives a unique informative code throughout the database. This should help researchers understand how to use the database in a responsible and informed way (the last section explains in more detail how we propose to deal with the different missingness scenarios). The coding and labelling protocols are as follows:

- *True lack of observed behaviour*: coded as **0**, no label since it represents 0 seconds or visits.
- *Error-induced lack of observed behaviour*: coded as **-2**, labelled as "Error-induced lack of behaviour".
- *Not enough information*: coded as **-1**, labelled as "Uncertain lack of behaviour".
- *Not re-contacted or refusal to participate in a given wave*: coded as **-3**, labelled as "NA: not in wave".
- *Not tracked*: coded as missing (i.e., ".").

Protocol for Naming Variables

All the online behavioural variables represent the average number of seconds or visits for which a participant did a specific behaviour, for a determined number of days.

The variable naming is structured in 4 to 6 different parts, depending on the variable in question:

- A prefix indicating whether the average was computed using the 15 days before participants started the survey, the 15 days after participants started the survey, or the entire tracking period: **PRE**, **POST** or **ALL**.
- Another prefix indicating whether the variable measures the average number of visits or seconds: **V** or **T**.
- For those variables measuring media exposure, another prefix which indicates whether the time or number of visits refers to all the types of articles, or articles about politics, national news, regional news, international news or opinion pieces: **Nw**, **Pol**, **Nat**, **Reg**, **Int** or **Op**. For those variables measuring the time or number of visits to a specific twitter profile, there is a fix suffix before the twitter handle: **TW.**
- The specific behaviour observed. This in general indicates the webpage and/or app in question (e.g., Facebook) or the group of webpages (e.g., Social media sites, or the entire Internet) observed. The lists of webpages and groups of webpages used by Netquest to construct each of the variables is provided here. The URLs included in each of those lists can be found here.
- A suffix, indicating the wave to which the variable belongs: **1**, **2** or **3**.

Taking all this into account, Table 4 displays some examples of variable names, also indicating their meaning and the group and wave to which they pertain.

Table 4 Examples of Online Behavioural Variable Names

Variable	**Meaning**	**Time frame**	**Measure**	**Type of media**	**Behaviour**	**Wave**
ALL_V_Facebook_1	Average number of visits to Facebook	All days of the wave	Visits	NA	Facebook	1
PRE_T_Nw_repubblica_2	Average time spent visiting news in “Repubblica”	The 15 days prior to the survey	Time	News	Repubblica	2
POST_T_Pol_50T_3	Average time spent visiting political news in the 50 most visited news outlets	The 15 days after the survey	Time	Politics	50 most visited news outlets	3
ALL_T_TW_Renzi_1	Average time spent visiting the twitter profile of Renzi	All days of the wave	Time	Twitter	Renzi’s profile	1

Source: own elaboration.

The database also includes some auxiliary variables, aimed at providing more information about the quality of the data. These fall in three categories:

1) **not_tracked**: this variable indicates whether an individual has their online behaviours tracked, or not.

2) **undercovered_device**: this variable indicates whether an individual has at least one device not covered, which could indicate that the observed behaviours might not be complete

3) **missclass_cat_*name of webpage*:** for Twitter, Facebook and the 10 most visited news media outlets in each country, we identified whether the observed behaviours where: (1) error-induced non-observations, (2) underestimated observations, (3) true non-observations, (4) true observations.

4) **Number of different types and devices tracked:** these variables give extra information about the number of devices and browsers that participants were tracked on, as well as the number of specific devices and browsers used by participants. These variables are the following: *desktop_windows_, desktop_apple_, desktop_windows_chrome_, desktop_windows_firefox_, desktop_apple_chrome_, desktop_apple_firefox_, desktop_apple_safari_, mobile_android_, mobile_ios_, num_browser_windows_tracked_, num_browser_apple_tracked_, num_devices_tracked_, num_mobile_tracked_, num_pc_tracked_.* All variables are continuous.

Protocol for Labelling Variables

Variable labelling seeks a balance between being informative and not being excessively long. Given that the variables' names all include information on the wave, this information is not repeated in the variables' labels. Thus, for any given variable available in different waves, all the variable labels are the same.

Protocol for Labelling Variable Values

Protocol of assignment of value labels to variables:

The assignment or not of value labels follows a precise protocol in the E-DEM dataset.

1. *If a variable includes non-response categories, it will at least have a generic value label to clarify the meaning of those responses* (i.e., to clarify that -1 means "Uncertain lack of behaviour"). This rule takes precedence over all the others, irrespective of the type of variable involved.
2. *Quantitative variables have no value labels* (except if they include non-response categories). This is the case for all online behavioural variables, since they all measure either number of visits or time spent in seconds.
3. *Auxiliary categorical variables are always labelled.* The previously presented auxiliary variables range from 2 to 4 categories per variable. They are always labelled.

The labels of the specific auxiliary categorical variables are always the same across waves:

- not_tracked (1= "Not tracked", 2= "Tracked")

- undercovered_device (1 = "Fully covered", 2 = "Undercovered")

- missclass_* (1 = “error-induced non-obs”, 2= “underestimated”, 3 = “correct non-obs”, 4 = “correct observation”).

- Number of different types and devices tracked: all are continuous

Naming and Labelling Language

Variable names, variable labels and value labels are all in English except when they refer to proper nouns, such as the names of media outlets (i.e., El Pais) and politicians (i.e., Pedro Sánchez) or the abbreviations of political parties’ names (i.e., UP, for Unidas Podemos), which are maintained in their original language.

Survey variables

6. Variable List

In this section, the complete list of online behavioural and methodological variables available in the integrated datasets is presented.

The list of variables is presented in tables, whereby the first column includes information on the variable names (when a variable is available in several waves, only the name of the first wave in which it appears is displayed), the second column displays the value label names (for all the variables that have value labels), the third column shows the variable labels (which clarify the contents of the variables), and columns four through seven inform of the wave or waves in which each variable is available (a capital “X” in a variable * wave cell indicates that the variable is available in the wave, and a blank space means that it is not).

To facilitate the navigation through the variable list, the information is presented in a series of tables, each of which referring to one group of variables: for the online behavioural variables, the table can be accessed in excel format here, given its share size. Table 5 presents the list of global variables. Table 6 presents the list of methodological variables.

Global Variables

Table 7 shows the list of global variables, which contain information on general characteristics of the survey. There are only two global variables, including the panelist' unique id number, the country of the panellist. and a variable indicating the days between the participant started the survey and it restarted after the treatment (only relevant for the first wave's experiment).

Table 5 List of Global Variables

Battery	Variable name	Value label	Variable label	W1	W2	W3
	g6	alpha	CodPanelista	X	X	X
	g8	country	SURVEYCOUNTRY	X	X	X
	g37	cont	Days past between the participant started the survey and restarted it after the treatment	X		
	wave_	wave	Participation in the wave		X	X

Source: own elaboration.

Online Behavioural variables

There are 1,693 online behavioural variables in the database, covering participants' general internet consumption, as well as their consumption of news and social media. All these variables were computed for each of the three waves, and they all are continuous.

Given the big volume of variables, the complete list for Italy can be accessed in excel format from here.

Methodological Variables

The databases also include some previously discussed methodological variables. The methodological variables were computed for waves 1 and 3. Table XX shows the complete list of variables.

Table 6 List of Methodological Variables

Battery	Variable name	Value label	Variable label	W1	W2	W3
	not_tracked_	not_tracked	Not tracked with a meter	X	X	X
	undercovered_device_	Undercovered_device	Whether the panellist has at least one device not tracked or not	X		X
	Missclass _FB_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _TW_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X

Battery	Variable name	Value label	Variable label	W1	W2	W3
	Missclass _repubblica_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _libero_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _corriere_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _gazzettadelsud_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _ansa_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _dagospia_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _lastampa_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _ilsole24ore_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _virgilio_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	Missclass _gds_	Misclass_*	Whether the observations for this domain are true or affected by errors	X		X
	desktop_windows_	cont	Number of desktop windows tracked	X		X
	desktop_apple_	cont	Number of desktop apple tracked	X		X
	desktop_windows_chrome_	cont	Number of chrome browsers tracked on windows	X		X
	desktop_windows_firefox_	cont	Number of firefox browsers tracked on windows	X		X
	desktop_apple_chrome_	cont	Number of chrome browsers tracked on apple desktop	X		X
	desktop_apple_firefox_	cont	Number of firefox browsers tracked on apple desktop	X		X
	desktop_apple_safari_	cont	Number of safari browsers tracked on apple desktop	X		X
	mobile_android_	cont	Number of android devices tracked	X		X
	mobile_ios_	cont	Number of iOS devices tracked	X		X
	num_browser_windows_tracked_	cont	Number of browsers tracked on windows	X		X
	num_browser_apple_tracked_	cont	Number of browser tracked on apple	X		X
	num_devices_tracked_	cont	Number of devices tracked	X		X
	num_mobile_tracked_	cont	Number of mobile devices tracked	X		X
	num_pc_tracked_	cont	Number of desktops tracked	X		X

Source: own elaboration.

7. Recommendations on How to Use the Database

Open access online behavioural databases are not common, and there is limited practical knowledge on how to better use them. This section provides some information on best practices that can be applied when using the TRI-POL online behavioural databases.

Dealing with the lack of observed online behaviours

Previously we described how we coded each type of lack of observed behaviour. Although much research is still needed to determine the best approach to deal with this lack of observed behaviours, we propose our approach to better deal with them.

For the specific variables for which we were able to collect enough auxiliary information (Facebook, Twitter and the TOP 10 most visited news outlet of each country), we can identify when a lack of observed behaviour was induced by errors or when it was real. For those, we propose the following:

1) *True lack of observed behaviours*: leave them as 0, which is their current value

2) *Error-induced non-observations*: re-code them as missing and exclude those participants from your analyses. Nonresponse weighting approaches can be used to re-adjust the sample. Proposed STATA 17 code:

```
foreach var of varlist ALL_V_Facebook_1-POST_V_SocialMedia_1{
recode `var' (-2 = .)
}
```

It is not mandatory for users to consider those participants as missing. We are ware that the information used to identify non-observations as error-induced comes from self-reported data, which can also be affected by errors. Therefore, some participants might as well be wrongly misclassified as missing when following our approach. Users can decide to consider those non-observations as real, being mindful that the decision will most likely inflate the measurement errors of their results. If someone wishes to do so, this is the code to use:

```
foreach var of varlist ALL_V_Facebook_1-POST_V_SocialMedia_1{
recode `var' (-2 = 0)
}
```

For all the other online behavioural variables, we did not have enough information to properly discern when a lack of observed behaviour for a participant affected by tracking undercoverage was real or induced by errors. Most past research, when faced with this, has considered these non-observations as reals, treating them as 0 seconds or 0 visits, in general. Nonetheless, we considered more transparent to give them a specific value and label, to quantify the level of uncertainty for each variable. Nonetheless, when conducting the analyses, researchers will need to decide whether they treat these non-observations as real (i.e., 0 seconds/visits) or induced by errors (i.e., missing). The proposed STATA 17 codes for both options are the following:

```
foreach var of varlist ALL_V_Facebook_1-
POST_V_SocialMedia_1{
recode `var'= (-1 = 0)
}
```

or

```
foreach var of varlist ALL_V_Facebook_1-
POST_V_SocialMedia_1{
recode `var'= (-1 = .)
}
```

Although we cannot provide advice on what to do with these non-observations, all the research published so far has treated them as real non-observations (i.e., 0 seconds/visits). If this is your choice, we recommend properly informing readers and reviewers about the proportion of participants which might have dubious non-observations treated as real.

Dealing with participants not in the specific survey wave

Previously we described how we coded participants who did not participate in the second and the third waves. For those we propose to simply treat them as missing:

```
foreach var of varlist ALL_V_Facebook_1-POST_V_SocialMedia_1{
recode `var' (-3 = .)
}
```

Dealing with participants with unrealistic behaviours

We recommend researchers using the TRI-POL databases to check for participants with unrealistic observed behaviours who might skew the distribution of their variables of interest.

Although theoretically possible, participants should not present values over 1,440 seconds (i.e., 24 hours) for the variables ALL/PRE/POST_T_ Internet_1/2/3, which measure the average time spent on the Internet. There are some reasons why this could, theoretically, be possible: 1) one individual might actively use more than one device at the same time, adding to more than 24 hours; 2) a device might be left unused with an active browser, while the participant uses other devices; 3) a participant might share one or more tracked device with other individuals. However, these cases should be a) rare and b) in most cases considered as of low quality. Technology malfunctions or non-human participants could be other possible explanations.

The proportion of participants above this threshold can be computed, for instance for wave 1, using the following code:

```
tab g6 if ALL_T_Internet_1 > 86400 & not_tracked == 2
```

We have not removed or coded these participants in any way, since there is still too little knowledge about what might cause these observations to happen, or the implications of keeping or excluding them from the analyses. However, we consider relevant for people to be mindful about their existence, and how they can affect their results.

Merging the behavioural and survey databases

The goal for most researchers using this database will be to combine it with the information about participant's attitudes, opinions and sociodemographic information collected through the different survey waves. Therefore, here we briefly explain how to merge both databases.

Both the survey and behavioural databases share the same ID variable, which is named g6 (labelled CodPanelista) in both databases. The following STATA 17 code can be used to merge both databases:

```
merge 1:1 g6 using “survey database file path”
```

8. References

Bosch, O.J., Revilla, M. (2022a) When survey science met web tracking: presenting an error framework for metered data. *Journal of the Royal Statistical Society: Series A* (Statistics in Society), DOI: 10.1111/rssa.12956

Bosch, O. J., & Revilla, M. (2022b). The challenges of using digital trace data to measure online behaviors: lessons from a study combining surveys and metered data to investigate affective polarization.In SAGE Research Methods Cases. https://dx.doi.org/10.4135/9781529603644

Bosch, O. J. (2022). Track me but not really: Tracking undercoverage in metered data collection. https://doi.org/10.31219/osf.io/2grpa