Understanding Society: *Calendar Year Dataset, 2020*

USER GUIDE

Version 1, *July 2022*

Understanding Society

Economic and Social Research Council

Contents

1. Introduction

The Calendar Year Dataset 2020 is designed to enable timely cross-sectional analysis of individuals and households relating to the situation in 2020. Such analysis can, however, only involve variables that are collected in every wave, as the data files combine data collected in each of three waves: analysis cannot be restricted to data collected in one wave during 2020, as this subset will not be representative of the population. This section provides an introduction to Understanding Society and to the structure of the Calendar Year Dataset 2020.

Understanding Society: the UK Household Longitudinal Study, started in 2009 with a general population sample of UK residents living in private households of around 26,000 households and an ethnic minority boost sample of 4,000 households. All members of these responding households and their descendants became part of the core sample who were eligible to be interviewed every year. Anyone who joined these households after this initial wave, were also interviewed as long as they lived with these core sample members to provide the household context. At each annual interview, some basic demographic information was collected about every household member, information about the household is collected from one household member, all 16+ year old household members are eligible for adult interviews, 10-15 year old household members are eligible for youth interviews, and some information is collected about 0-9 year olds from their parents or guardians. Since 1991 until 2008/9 a similar survey, the British Household Panel Survey, was fielded. The surviving members of this survey sample were incorporated into Understanding Society in 2010. In 2015, an immigrant and ethnic minority boost sample of around 2,500 households was added. To know more about the sample design, following rules, interview modes, incentives, consent, questionnaire content please see the study overview and user guide.

Each set of annual interviews are referred to as a wave. The fieldwork period for interviews of each wave stretches over 24 months. But the time interval between two consecutive wave interviews for each person and household is generally around one year. The way this is operationalised is by having overlapping waves, see Figure 1. Sometimes individuals are difficult to contact or are away during the entire intended fieldwork period. In such cases, interviews are scheduled in the weeks after the intended fieldwork period has ended. As a result, even though the intended fieldwork period for any wave is 24 months, the actual fieldwork may be extended by a few months. So, interviews for Wave 1 stretched from January 2009 to March 2011. But only a very small proportion of interviews for any wave are conducted in the 3rd year, less than 5%.

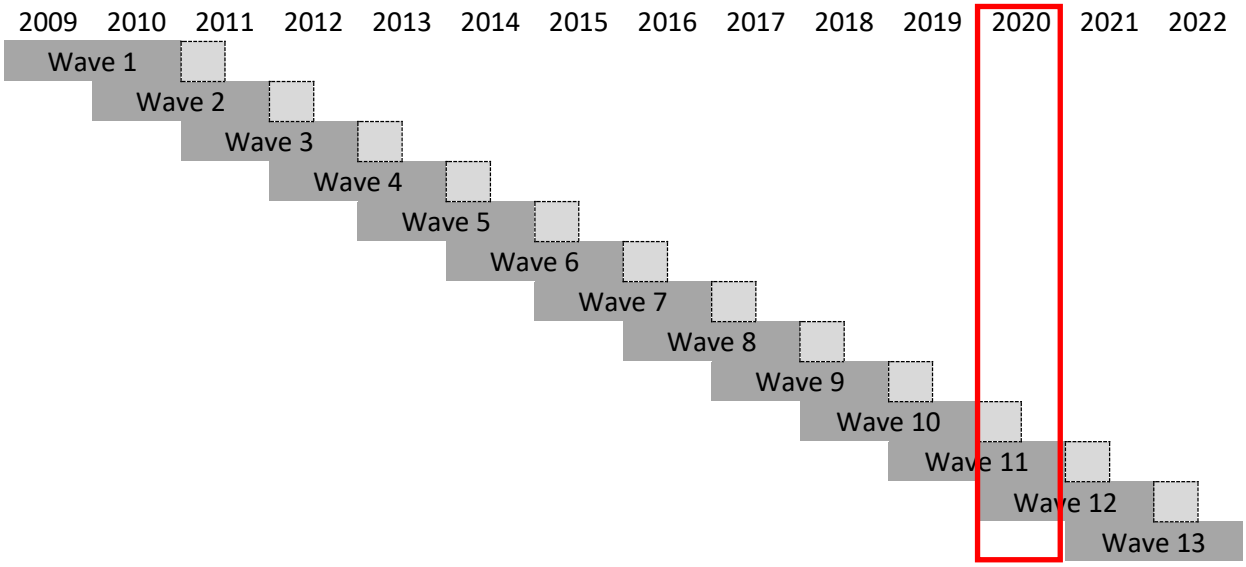

Figure 1: Fieldwork period

The 24-month fieldwork period is applicable for these samples only: GPS-GB part and EMBS. Interviews for IEMBS take place in Year 2 of each wave fieldwork period and interviews for BHPS and GPS-NI samples takes place in Year 1 of each wave fieldwork period. See Figure 2.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
GPS – GB part, Year 1 sample	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
GPS – GB part Year 2 sample		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13
GPS – NI part	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
BHPS		W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
EMBS Year 1 sample	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
EMBS Year 2 sample		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13
IEMBS		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13

Figure 2: Fieldwork period by samples

To analyse change over time information needs to be collected at fixed intervals for all sample members. This is the case for Understanding Society, where even though there is overlapping fieldwork period for some samples, the information collected for any sample member across two consecutive waves is always around one year apart. However, to analyse information collected in a specific year will require combining information collected in that year as part of consecutive waves. For example, to analyse information collected in 2020, we will need to combine data collected as part of Waves 10, 11 and 12.

Secondly, the survey data is standardly provided as wave specific files and so anyone interested in doing this type of analysis would have to combine the data from across three consecutive waves that was collected in the year of interest. So, to analyse data collected in 2020, analysts would have to combine datafiles for Waves 10, 11 and 12.

Finally, as data for wave n is collected in year t, t+1 and t+2 and released in November of year t+2, anyone interested in analysing data for year t+1 would have to wait until November of year t+2. For example, to analyse data collected in 2020, data analysts will need data collected in Waves 10, 11 and 12 (see Figure 1), but data for Wave 12 will be released in November 2022. *Please note that as of November 2021, data from Waves 1-11 have been released and so data analysts will be able to produce calendar year datasets for years 2009-2019 by themselves.*

To enable cross-sectional analysis relating to a calendar year to be carried out more easily and in a timely manner than has been the case until now, we have decided to combine data collected in a specific year from across multiple waves and release it as separate calendar year datasets, with appropriate analysis weights, starting with *this* 2020 Calendar Year dataset. In each subsequent year, additional yearly datasets will be released, and we will aim to release these datasets in December of the following year. So, the 2021 dataset will be released in December 2022 and so on.

There are two versions of the 2020 Calendar Year dataset. The End User Licence (EUL) version (SN 8988) is suitable for the majority of research, however, a Special Licence (SL) version (SN 8987) is also available which contains more detailed variants of some variables, non-top-coded income variables plus additional derived variables. To identify the variables in the SL version but not in EUL version please refer to this document . Please note that access to the Special Licence dataset requires the completion of an application form and the associated processing time is greater than to access the EUL version. For more details please refer to the information on the Access Understanding Society data page which refers to the access conditions outlined by the UK Data Service.

2. Data structure

2.1 Who is included?

The annual release files include all households where the first individual adult interview per household takes place within the calendar year (2020). This approach has been taken to preserve the integrity of a household and provide consistency for future calendar releases. This does mean that at the end of the year boundary, we do include some individuals who were interviewed early in 2021 as at least one member in that household responded before the end of 2020 (98.3% of the individual interviews included were completed in 2020). These households and their members will therefore not be included in next year's calendar release file. By the same token, some individual interviews carried out in early 2020 will not be included in the calendar year 2020 file, as at least one other household member completed their interview in 2019.

As interview dates for two consecutive waves are not exactly one year apart for everyone, there are a few individuals who were interviewed twice in 2020. Around 91.5% of adult respondents were interviewed once in 2020, 1.2% were interviewed once in 2021 and 7.3% were interviewed twice (2.8% were interviewed as part of Waves 10 and 11, and 4.5% were interviewed as part of Waves 11 and 12). See Appendix 1 for further details. The data consist of wave 10 data for 523 households

(924 adult interviews), wave 11 data for 7,570 households (13,577 adult interviews) and wave 12 data for 9,848 households (17,346 adult interviews). *Note that these multiple observations from the same individuals must be retained in analysis to provide a representative sample: do not exclude apparent duplicates.*

2.2 Datafiles

The purpose of the 2020 Calendar Year dataset is focused on households where interviews were obtained. So, while most datafiles from the main survey annual data release are also included in the 2020 Calendar Year dataset, files that mostly include variables that relate to households where an interview was not obtained, **callrec**, **hhsamp, indsamp** are excluded.

As the Calendar Year dataset is to be used for cross-sectional analysis, this relies on the inclusion of both year 1 and year 2 sample members (see section 3.1). Thus, the 2020 dataset consists largely of Wave 12 interviews with the year 1 sample and Wave 11 interviews with the year 2 sample. Therefore, any variables that are only collected in some of the waves included in this dataset are also excluded. As the files **chmain** (Wave 11 only) **and parstyle** (Wave 12 only) consist entirely of such variables these are excluded as well. *The only exceptions to this rule are the wave specific household identifiers and the politics and election variables.* The household identifiers are included as these are wave specific (there is no concept of a longitudinal household) and needed for identifying individuals in the same household. While some of politics questions are asked every wave, there were additional questions about the General Elections, the EU Elections, the EU referendum, as well as general questions on political engagement, political efficacy, social and political values and views on EU membership that were triggered by the General Elections or asked only in Wave 12 but needed for elections related analysis (see Appendix 2 for a list of these variables). *Please note, any estimates based on questions asked in Wave 12 only in this 2020 calendar year dataset will be biased as these will be based on Year 1 sample which is not representative of the population. Please refer to the main user guide, particularly the weighting FAQ Question Number 11 where this issue is explained.*

Questionnaires for all waves including Waves 10, 11 and 12 are available in the Study documentation.

The files that are included are shown in Table 1. The root names for the files and variables in the main survey annual data release have been retained, but with a different prefix (for details see Section 2.4).

Table 1: List of files in the 2020 Calendar Year dataset

Filename	Description
wxy_indall	Household grid data for all persons in household, including children and non-respondents
wxy_hhresp	Substantive data collected from responding households
wxy_indresp	Substantive data collected from responding adults (16+) including proxies. Some information collected in these questionnaires are better presented in multi-level files (see Table 2).

wxy_youth	Substantive data from youth questionnaire
wxy_child	Childcare, consents and school information of all children (0-15 years) in the household. This is a derived data file collecting information pertaining to children as reported by their parents and guardians in the adult questionnaire.
wxy_egoalt	Kin and other relationships between pairs of individuals in the household. This is a derived data file based on information collected in the household grid about relationships between household members.
wxy_income	This file contains reports of unearned income and state benefits for each individual.
wxy_newborn	Every wave after Wave 1, basic information about newborn children such as brithweight, etc. is collected from new parents

2.3 Missing values

Missing values are flagged in the same way as the main survey annual data release (see the missing values). An additional missing value code, -14, is available in this dataset. This code indicates a variable was not present in that wave. As variables not collected in all three waves except for the wave specific household identifiers and politics and election related variables are excluded, this missing value code only applies to those variables. This is an additional code and not to be confused with -8 (inapplicable) which indicates the variable was asked in the wave but not asked of the respondent due to questionnaire routing.

2.4 Naming conventions

Filenames

Files are prefixed with a wave identifier indicating the waves contained within the file. For the 2020 calendar release, the file prefix is **jkl_** as data was obtained from households in Waves 10, 11 12 as shown in Figure 1, as the wave prefixes for these waves are **j**, **k** and **l**.

Variables

Variable names, like filenames are prefixed with a wave identifier indicating the waves contained within the file, and so for the 2020 calendar release, the variable names are prefixed with **jkl_**. **period.** For example, the variable **jkl_country** refers to the country the sample members are living in 2020 with this variable being a combination of **j_country**, **k_country** and **l_country**. Variables like **pidp**, **ethn_dv** which do not change across waves do not have a prefix. Another set of variables without the **wxy_** prefix are the three wave specific household Identifiers: **j_hidp**, **k_hidp** and **l_hidp**. These record the household identifier of the wave where each record originates.

2.5 Identifiers

The variable ***wxy*_wave** records the interview wave that each observation is taken from. In the 2020 dataset, around 55% of the observations in **jkl_indresp** (adult respondents) are from Wave 12, 42% from Wave 11, and 3% from Wave 10.

The variable **pidp**, which is the unique cross-wave person identifier, for every sample member is also provided. But, as a few individuals can be interviewed in the same year as part of two consecutive waves each data record (or row in data file) is *not* uniquely identified by **pidp**, rather it is the combination of **pidp** and ***wxy*_wave** that uniquely identifies each data record.

Households are uniquely identified in each wave by ***w*_hidp**, a wave specific variable with a different prefix for each wave. It can be used to link information about a household from different records **within a wave** but cannot be used to link information **across waves**. Since the composition of households can change between waves, **the data do not include a longitudinal household identifier**. For example, **j_hhresp** can be linked with **j_indresp** using **j_hidp** but not with **k_hhresp**. In the calendar file, we have included the ***w*_hidp** variables for the three waves making up each dataset, i.e., the variables ***w*_hidp**, ***x*_hidp** and ***y*_hidp.** We have combined these into the variable ***wxy*_hidp** which contains the **hidp** of the wave for each data record. As the **hidp** variables for the three waves are mutually exclusive (i.e., there are no overlaps), ***wxy*_hidp** uniquely defines each data record in household level files ***wxy*_hhresp**.

2.6 Key variables

Table 2 shows a list of key variables that are available in this data and the files in which these are available. You can also link this data to the main annul release datasets. See Section 2.7 for further details on how to do this.

Table 2: Key variables

Topic domain	Variable name	Short description	Datafiles
Identifiers			
	pidp	unique cross-wave person identifier	All individual files
	*wxy*_hidp	household identifier	All files
	*w*_hidp, *x*_hidp and *y*_hidp	wave specific household identifier that equals the household identifier for the wave from which the record originates, -14 otherwise	All files
	*wxy*_pno	person number in the household identified by *wxy*_hidp	All individual files
	*wxy*_mnpno	PNO of biological mother	indall, indresp, child, youth
	*wxy*_fnpno	PNO of biological father	indall, indresp, child, youth
	*wxy*_mnpid	PIDP of biological mother	indall, indresp, child, youth
	*wxy*_fnpid	PIDP of biological father	indall, indresp, child, youth
	*wxy*_mnspno	PNO of biological, step or adopted mother	indall, indresp, child, youth
	*wxy*_fnspno	PNO of biological, step or adopted father	indall, indresp, child, youth
	*wxy*_mnspid	PIDP of biological, step or adopted mother	indall, indresp, child, youth
	*wxy*_fnspid	PIDP of biological, step or adopted father	indall, indresp, child, youth
	*wxy*_grmpno	PNO of grandmother	indall, indresp, child, youth
	*wxy*_grfpno	PNO of grandfather	indall, indresp, child, youth
	*wxy*_childpno	PNO of child	indresp, child
	*wxy*_ppid	PIDP of partner	indall, indresp
	*wxy*_ppno	PNO of partner	indall, indresp
	*wxy*_sppid	PIDP of spouse	indall, indresp

	*wxy*_sppno	PNO of spouse	indall, indresp
Residential location			
	*wxy*_country	Country in the UK sample members currently live in	hhresp , indall, indresp, child, youth
	*wxy*_gor_dv	Region in the UK	hhresp , indall, indresp, child, youth
	*wxy*_urban_dv	Urban or rural area, derived	hhresp , indall, indresp, child, youth
Demographic characteristics			
	*wxy*_sex_dv	sex, derived	indall, indresp, child, youth
	*wxy*_doby_dv	Year of birth, derived	indall, indresp, child, youth
	*wxy*_age_dv	age at time of interview, derived	indall, indresp, child, youth
	*wxy*_mastat_dv	marital status	indall, indresp
	*wxy*_nchild_dv	number of children in the household. Includes natural children, adopted children and stepchildren, under the age of 16	indall, indresp
	*wxy*_jbstat	employment status	indresp
	wxy _ethn_dv	ethnic group - derived from multiple sources	indall, indresp, youth
	bornuk_dv	Whether born in the UK or not	
Socio-economic characteristics			
	*wxy*_hiqual_dv	highest qualification status	indresp
	*wxy*_jbsoc00_cc	Standard Socio-economic Classification (SOC 2000) of current job. Condensed three-digit version status	indresp
	*wxy*_jbnssec8_dv	current job: Eight Class NS-SEC status	indresp
	*wxy*_jbnssec5_dv	current job: Five Class NS-SEC status	indresp
	*wxy*_jbnssec3_dv	current job: Three Class NS-SEC status	indresp
	*wxy*_fimnnet_dv	own total estimated net monthly income status	indresp

	*wxy*_fimnlabnet_dv	own total estimated net monthly income from labour status	indresp
Health			
	*wxy*_sf12mcs_dv	SF-12: mental health component score, derived status	indresp
	*wxy*_sf12pcs_dv	SF-12: physical health component score, derived status	indresp
	*wxy*_health	long-standing illness or disability status	indresp
	*wxy*_scghq1_dv	subjective wellbeing (GHQ): Likert status	indresp
	*wxy*_scghq2_dv	subjective wellbeing (GHQ): Caseness status	indresp
Household-level characteristics			
	*wxy*_hhsize	number of individuals in the household	hhresp, indall, indresp
	*wxy*_nkids_dv	number of children aged under 16 in the household	hhresp, indall, indresp
	*wxy*_hhtype_dv	household composition	hhresp, indall, indresp
	*wxy*_tenure_dv	housing tenure	hhresp
	*wxy*_fihhmnnet1_dv	net household monthly income	hhresp
	*wxy*_ieqmoecd_dv	household income conversion factor (modified OECD scale)	hhresp

2.7 Linking datafiles

To link different files within the calendar year dataset

To link different individual level files use **pidp *wxy*_wave**. Here is an example Stata code to link **jkl_indresp** with **jkl_indall**

```
use jkl_indall, clear
merge 1:1 pidp jkl_wave using jkl_indresp
```

To link individual and household level files use ***wxy*_hidp**. Here is an example Stata code to link **jkl_hhresp** with **jkl_indall**

```
use jkl_indall, clear
merge m:1 jkl_hidp using jkl_hhresp
```

To link files in the calendar year dataset with the main annual release datafiles

Only individual level files can be linked using **pidp**. But as **pidp** does not uniquely identify each row in the ***wxy_*** individual level datasets the merging command will need to specify that. Here is an example Stata code to link **jkl_indresp** with **xwavedat**, and **jkl_indall** with **a_indall**

```
use jkl_indresp, clear
merge m:1 pidp using xwavedat

use jkl_indall, clear
merge m:1 pidp using a_indall
```

2.8 Geographical data linkage

The standard datafiles available with this 2020 Calendar Year dataset includes Government Office Regions (GOR) as the lowest level geography available. Other more detailed (or lower level) geographical identifiers with LSOA being the lowest available are released as part of the main annual survey data and are available as Special Licence datasets. For a full list of available geographies see https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000053#!/access-data. These datasets contain a file for each wave containing the geographical identifier and the wave specific household identifier (***w*_hidp**). As the 2020 Calendar Year dataset includes data from Waves 10, 11 and 12 and includes the wave specific household identifiers**, j_hidp k_hidp l_hidp**, the files can be matched with the Wave 10-12 geography files using these household identifiers.

3. Analysis guidance

These calendar year datasets should be used for cross-sectional analysis only. For those interested in longitudinal analyses using Understanding Society please access the main survey datasets from the UKDS: End User Licence version or Special Licence version.

3.1 Weighting, clustering, stratification and representativeness

Users should always use the clustering variable (**psu**), stratification variable (**strata**) and a weight when analysing data from Understanding Society data. Only cross-sectional weights are provided for this release.

The weights for this release are calculated using the same methods as those used in the main data release with the exception that the nonresponse model is run separately for each of the earlier waves (**w** and **x**) and is run only for the sample issue months 1-12 for the most recent wave (**y**).

The combined sample represents the full population. This is because it contains the correct balance of year 1 and year 2 sample members and of prompt and late respondents. The year 1 and year 2 samples are rather different in structure and not representative unless combined. Similarly, late respondents (those issued in one year but not interviewed until the following year) are likely to have distinct characteristics. In the calendar year data file, late respondents of the most recent wave (**y**), that fall outside of the calendar year, are compensated by late respondents of the earliest wave (**w**), where only these are included in the dataset. For example, in calendar year 2020, late non-respondents of Wave 11 (year 2 sample) that completed the questionnaire in 2021 are excluded from the dataset. But these are compensated by late Wave 10 (year 2 sample) respondents who were supposed to complete interviews in 2019 but did not do so until 2020. Assuming that late responders have broadly similar characteristics each year, this allows an even representation of all types of respondents and therefore different groups of the population leading to a full representation of the population.

3.2 Income variables

Information about the income variables can be found in the Understanding Society Main Study documentation.

3.3 Main Study changes due to the COVID-19 pandemic

Due to the Covid-19 pandemic, face-to-face interviewing was suspended from April 2020 and eligible sample members were invited to complete the questionnaires online, with non-responders followed up by interviewers for a telephone interview. Some questions were also introduced to the main survey in response to the pandemic such as questions about experiencing symptoms or Covid-19 or being diagnosed with it and experiences with the new furlough scheme that was introduced. At this time Wave 11 and 12 interviews were being fielded and so these changes affected these interviews. All these changes in fieldwork and mode during 2020 and its impact on analysis and non-response have been documented in Understanding Society Main Study changes due to the COVID-19 pandemic (Wave 11 release) as part of the Main Study User Guide. The document sets out the

changes to the fieldwork, the questionnaire (including new questions) the impact on response rates and derived variables. Similar changes were introduced for the Wave 12 interviews during 2020. Also see "COVID-19 and mode selection effects in Understanding Society" to know more about the mode changes during the pandemic on response rates.

4. Data access and citation

4.1 Citing this data

The citation changes at each release to reflect the addition of the data from the new wave. Please visit https://www.understandingsociety.ac.uk/documentation/citation for the citation for the latest version of the data. Search for "The bibliographic citation for the Understanding Society Calendar Year data collection"

Please cite each dataset that you use.

If you use Understanding Society data you must acknowledge this.

All works which use or refer to these materials should acknowledge these sources by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations must appear in footnotes or in the reference section of publications.

4.2 Citing this User Guide

When citing this User Guide, you can use the citation of this particular version quoted below. Note that where an online version is available on the Understanding Society website it is always the most up to date.

Institute for Social and Economic Research. (2022), *Understanding Society: Calendar Year Dataset, 2020, User Guide, Version 1.0, July 2022*, Colchester: University of Essex.

5. Help and support

5.1 User Guide and online documentation

Information about Understanding Society main survey, including the user guide, questionnaires, variables search, data management syntax files, data access information and so on can be found in the Study documentation.

5.2 Training, FAQ, Videos

The Help and Support section of the website provides links to the FAQ, online training courses and upcoming in-person training workshops. Training videos and webinars are available on our YouTube channel.

5.3 User Support

Questions about the data can be posted on our User Support Forum. Questions asked by other data users are also visible and searchable. Questions about the data and requests for one-on-one help sessions with user support team members can also be emailed to User Support at usersupport@understandingsociety.ac.uk.

5.4 Publications Library

To see an up-to-date list of research publications using Understanding Society data, visit the Understanding Society website: https://www.understandingsociety.ac.uk/research/publications.

Appendix 1: Adult interview dates & waves

Interviewed in 2020	98.3%
Interviewed in 2021	1.7%
Interviewed once in 2020	91.5%
Interviewed once in 2021	1.2%
Interviewed twice (Waves 10 & 11)	2.8%
Interviewed twice (Waves 11 & 12)	4.5%
Interviewed once in 2020	91.5%
As part of Wave 10	1.5%
As part of Wave 11	38.4%
As part of Wave 12	51.6%
Interviewed twice in 2020	6.4%
As part of Waves 10 & 11	2.5%
As part of Waves 11 & 12	3.9%
Interviewed once in 2020 & once in 2021	0.9%
As part of Waves 10 & 11	0.3%
As part of Waves 11 & 12	0.6%
Interviewed once in 2021	1.2%
As part of Waves 11	0.6%
As part of Waves 12	0.6%

Appendix 2: Political and Elections questions

Variable names	Wave 10	Wave 11	Wave 12
vote1 – vote8	√	√	√
euparl	√	√	√
voteeuparl	√	√	√
eumem	√	√	√
votereas_coded (General Elections triggered)		√	√
voteimp_coded (General Elections triggered)		√	√
votetxspnd (General Elections triggered)		√	√
voteeuint (General Elections triggered)		√	√
colbens1-6 (General Elections triggered)			√
perpolinf			√
civicduty			√
polcost			√
votenorm			√
voteintent			√
perbfts			√
grpbfts			√
demorient			√
poleff1-poleff4			√
scwhorupol			√
orgm1, orgmcawi1, orga1, orgmt1, orgat1			√
swvt1, swvt1w, wswvt2, wswvt3, wswvt2w, wswvt3w, wnivt1, wnivt2, wswvt2w_all, wswvt3w_all, wswvt2_all, wswvt3_all, wnivt2_all			√
opsoca-opsocf opsock-opsocp			√
immecon			√
immcultur			√