Understanding Society

Main Survey User Guide

Last revised 30 October 2024

Contents

Downloading Understanding Society

Download the data

Understanding Society data consists of multiple files, downloadable as a single zipped folder and released through the UK Data Service, in SPSS, Stata and tab delimited ASCII formats.

The access principles and guidance outlined in this section are derived from the Understanding Society Data Access Strategy.

Every November data from Wave 1 until the latest Wave are released. Data from Wave 1 - Wave 14 were released in November 2024 and in November 2025 data from Wave 1 - Wave 15 will be released. Data from previous waves are released each time, so that any corrections can be incorporated. These changes are included in a revisions file. Along with the data files, the zipped folder will contain this Main User Guide at the time of release, other User Guides, the revisions file, the questionnaires etc. However we recommend using the Understanding Society webpages as the documentation will develop over time, with the latest versions updated to these pages.

Researchers who would like to use Understanding Society need to register with the UK Data Service before being allowed to download the dataset. Researchers should also read the UK Data Service Research Data Handling and Security: Guide for Users before downloading the data. Under the terms and conditions of access this is required reading for users who obtain Special Licence data or Secure Access data, where disclosure risk is increased.

The majority of users will need the End user Licence (EUL) data which requires users to agree to the licence agreement online after which they will be given instant access to the data. Datasets which fall under this access level category include most of the data and only exclude information that could be disclosive, such as day and month of birth, detailed country of birth, etc. This document describes the differences between the Special Licence (SN 6931) and the End User Licence (SN 6614) versions. Files containing levels of geography lower than regions (previously Government Office Regions) are also excluded. Geographical indicators and other linked data available with the survey data are explained on our Linked Data page.

Download the EUL version of Understanding Society data direct from the UK Data Service. This video outlines how to download the Understanding Society EUL Data. Visit our Data Releases page to see the latest data available and when further updates are expected.

Access levels explained

There are different levels of access, depending on what information you want to use. Users are required to sign licence agreements, details of which are listed on the Access Understanding Society data page.

Once access has been approved, data can be downloaded from the UK Data Service.

End User Licence data (Safeguarded)

In preparing the data for the general release we have taken steps to maintain the confidentiality of responses. These include not releasing the full date of birth and not releasing the most detailed job-related SOC and SIC codes. Information on income and investment has been top coded. Open or narrative text, e.g., names of schools or employers, has not been released since it may indirectly identify individuals. Geographical identifiers below the level of GORs are also not included in the EUL release. This type of dataset should be suitable for the majority of researchers. The full set of Understanding Society EUL datasets available at the UK Data Service are identified in GN 33423.

Special Licence data (Safeguarded)

A number of sensitive data are released under Special Licence. Researchers can apply for access to Special Licence data through the UK Data Service if they can justify their research objectives and clearly explain why Special Licence data are needed for their project. It should be noted that some Special Licence datasets have restrictions on which countries researchers may apply for them. These restrictions can be identified in the ‘Access data’ tab on the UK Data Service catalogue webpage of any dataset. Further details are available on the Access Understanding Society data page. The full set of Understanding Society Special Licence datasets available at the UK Data Service are identified under GN 33428.

Secure Access data (Controlled)

Some data which is more disclosive than Special License versions can be only accessed by approved researchers in secure settings. To apply for Secure Access data through the UK Data Service researchers must be based at a UK academic institution or an ESRC-funded research centre. Applications may also be made through the Office for National Statistics Secure Research Service (ONS SRS). However, data from the ONS SRS do not contain all of the variables in the equivalent study available from the UK Data Service; the variables containing organisational names have been removed at the ONS SRS. Please check the data dictionaries to confirm whether the variables you require are available. In all cases Secure Access data applicants must be ESRC accredited researchers. Please contact the UK Data Service or ONS SRS for details of how to become an accredited researcher. Further details are available on the Access Understanding Society data page.

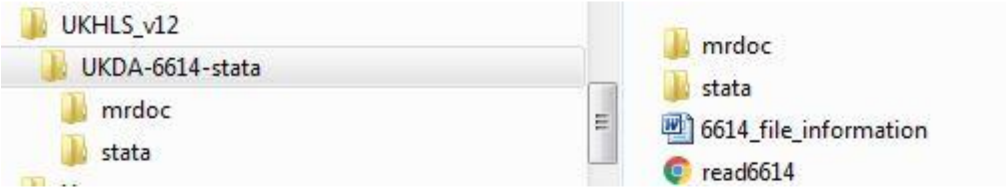
A few Understanding Society datasets fall into this category: SN 6676 which includes postcode grid references (but not postcodes themselves) and full date of birth and other linked administrative linked datasets. These datasets at the UK Data Service are identified under GN 33429. For access at the ONS SRS please contact them directly.

Downloaded data folder explained

Data are available in Stata, SPSS and Tab delimited ASCII format. Once you have chosen the data format, you will be asked to download a zipped folder.

To extract the data use software such as 7-ZIP. In this example we are using the Stata version named **UKDA-6641-stata**. Open the zipped file and click on Extract in the task bar, then choose a location to place the files. We recommend you keep all files in a designated folder, such as **UKHLS** and include the release version in the folder name as our data gets updated over time.

Once the files are extracted you will find two main folders and two files. The **read6614.htm** file provides basic information about this release and the word file **6614_file_information** a list of all files.



Documentation

The sub-folder **mrdoc** contains detailed study documentation such as user guides, questionnaires, fieldwork materials for each wave and a data dictionary for each data file.

Data files

The Stata data files are available in Stata 13. The sub-folder **stata\stata13_se** contains the data in study specific sub-folders: folder **bhps** contains BHPS files from all waves (most of which are harmonized) and the non-harmonized BHPS cross-wave files (**xwlsten**, **xwaveid_bh**) and the folder **ukhls** contains data files from all waves of UKHLS and harmonized and non-harmonised UKHLS cross-wave files (**xwaveid**, **xwavedat**). Note while **xwaveid** only includes information from UKHLS waves and **xwaveid_bh** only information from BHPS Waves 1-18, **xwavedat** is harmonised across BHPS and UKHLS and includes time-invariant information about all BHPS and UKHLS sample members collected across BHPS Waves 1-18 and all UKHLS waves currently available. The structure is the same for SPSS and tab delimited files. SPSS is available in version 25 under sub-folder **spss\spss25**.

To explore the main Study data structure and documentation take a look at the Study overview. To understand the file naming conventions, their meaning and missing values visit the Data section.

Tips for analysts:

The Understanding Society dataset is large and complex. If you are new to using these kinds of datasets you may benefit from our Help for new users page which outlines a pathway to exploring the data and resources. Our training courses enable you to learn how to put together the data files for different types of analysis and you can also take a look at our Cheat sheets – which are bare bones syntax files showing you how to perform common data management tasks.

For specific queries on using Understanding Society, please contact our User Support.

Ethics

Collecting, using and sharing data in research with people requires that ethical and legal obligations are respected.

The Understanding Society study protocols and research programme are scrutinised by a number of research ethics committees to assure that ethical and legal obligations are respected at all times.

Ethical approval statement

The University of Essex Ethics Committee has approved all data collection on Understanding Society main Study and innovation panel waves, including asking consent for all data linkages except to health records. Requesting consent for health record linkage was approved at Wave 1 by the National Research Ethics Service (NRES) Oxfordshire REC A (08/H0604/124), at BHPS Wave 18 by the NRES Royal Free Hospital & Medical School (08/H0720/60) and at Wave 4 by NRES Southampton REC A (11/SC/0274). Approval for the collection of biosocial data by trained nurses in Waves 2 and 3 of the main survey was obtained from the National Research Ethics Service (Understanding Society - UK Household Longitudinal Study: A Biosocial Component, Oxfordshire A REC, Reference: 10/H0604/2).

For further details on the various committees which have provided ethical approval of the Understanding Society Study and its components as appropriate see below:

Main survey: Ethics approval was received from the University of Essex Ethics Committee

- By letter dated 6 July 2007 for Waves 1 and 2
- By letter dated 17 December 2010 for Waves 3 to 5
- By letter dated 20 August 2013 for Waves 6 to 8
- By letter dated 4 October 2016 for Waves 9-11
- Ethics Approval number ETH1920-0123 for Wave 12
- Ethics Approval number ETH2021-0015 for Wave 13
- Ethics Approval number ETH2122-0246 for Wave 14
- Ethics Approval number ETH2223-0264 for Wave 15
- Ethics approval number 22/EE/0260 was received from the NHS Research Ethics Committee for Wave 16

Linkage to health records

- National Research Ethics Service (NRES) Oxfordshire REC A (08/H0604/124): 21 October 2008
- NRES Royal Free Hospital & Medical School (08/H0720/60): 18 June 2008
- NRES Southampton REC A (11/SC/0274): 28 September 2011 and 24 November 2011

Health Assessment and IBIO pilot

- NRES Oxfordshire REC A (10/H0604/2): 9 April 2010
- NRES Oxfordshire REC A (10/H0604/62): 19 August 2010
- NRES Oxfordshire REC A (10/H0604/70): 20 January 2011

Citation

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. 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.

Citing this User Guide

When citing this User Guide you can use the citation of this particular version quoted below. Alternatively, you can cite a previous version if required by replacing the date with that on the front of the User Guide you wish to cite. Note that the version available on the Understanding Society website is always the most up to date.

Institute for Social and Economic Research. (2024). *Understanding Society: Waves 1-14, 2009-2023 and Harmonised BHPS: Waves 1-18, 1991-2009, User Guide, 30 October 2024*, Colchester: University of Essex.

Study overview

Understanding Society is a longitudinal household panel study. The Study started in 2009 and follows on from the British Household Panel Study which ran from 1991-2008. Taken together the two Studies currently provide researchers with data on households in the UK spanning over 30 years.

As an introduction to the Understanding Society main Study data and documentation we particularly recommend the Help for new users pathway. This provides new users with a pathway to explore the Study and highlights the ways in which you can use the online resources to help you start using the data. These include:

- Training videos playlist Exploring Understanding Society, Data Structure and Selecting Weights on our YouTube channel.
- Variable descriptions and search facility. Find the variables you need for your research by searching by variable name, by data file or by index term. Explore How to use the Variable Search. This facility provides links between questions, variables and data files. Our list of key variables for the analysis of individual response data may also help you.
- The Code creator extracts data from the Main Understanding Society dataset and provides ready-to-use syntax to run on the downloaded data. Select the variables needed from the Variable Search, 'save' the variables and 'build' the code.
- Index Terms cover all the thematic areas in the Study. Use the Index terms to identify the variables most relevant to your research interests and to find other variables with related data throughout the dataset.
- What's new and notes on variable naming conventions. These tell you what's new for the latest Wave and how the data in each wave are named.
- Example syntax files show you how to perform common data management tasks like matching household and individual data files, producing long format individual level file for a number of waves, etc.
- Our training courses and user support for users. If you're starting out, you can ask to speak with someone from our team directly via our online helpdesk. If you want to examine more complex data questions you can visit our User Support Forum.

Read more About the Study and our Funders.

Study design

Understanding Society is a panel survey of UK households with yearly interviews.

The Study began in the UK in 2009-10. The overall Study has multiple sample components to enable research of different sub-groups over time and location or geography:

- 2009-2010: The **General Population Sample (GPS):** (i) a clustered and stratified, probability sample of approx. 24,000 households living in Great Britain in 2009-10 (ii) a simple random sample of approx. 2,000 households living in Northern Ireland in 2009 (selected with twice the selection probability as the Great Britain part). See Sample Design Paper.

- 2009-2010: The **Ethnic Minority Boost Sample (EMBS):** approx. 4,000 households selected from areas of high ethnic minority concentration in 2009-10 where at least one member was from an ethnic minority group. See Screening questions Appendix III and sample design paper - Design of the Understanding Society ethnic minority boost sample.

- 2010: The **British Household Panel Survey sample (BHPS)**, added in Wave 2: approx. 8,000 households from the BHPS sample in 2010.

- 2015: The **Immigrant and Ethnic Minority Boost Sample (IEMBS)**, added in Wave 6: approx. 2,900 households selected from areas of high ethnic minority concentration in 2015 where at least one member was born outside the UK, or from an ethnic minority group. See Screening questions Appendix III and sample design paper - Design and implementation of a high quality probability sample of immigrants and ethnic minorities: lessons learnt.

- 2022-2023: A **General Population Sample boost (GPS2):** added in Wave 14: approx., 5,700 households. A clustered in Great Britain but unclustered in Northern Ireland sample (selected with equal selection probability as the Great Britain part) in 2022-2023. See Wave 14 Boost technical report.

The What are the different samples video gives an overview of the sample design for Understanding Society and how to account for it in analysis. Note this video does not mention the latest sample added, GPS2.

Tips for analysts:

Analyse all samples that are available together. As the data for all samples are provided in the same data file, as long as you don't explicitly exclude a sample you will be doing this by default. You can analyse the GPS or the GPS2 by themselves, but it will mean your sample size will be smaller than if you were using all samples and the sample sizes will not be large enough for ethnic minority and immigrant groups. Additionally there are no appropriate weights to use any of the samples by themselves, and so the estimates based on analysing these samples by themselves may be biased.

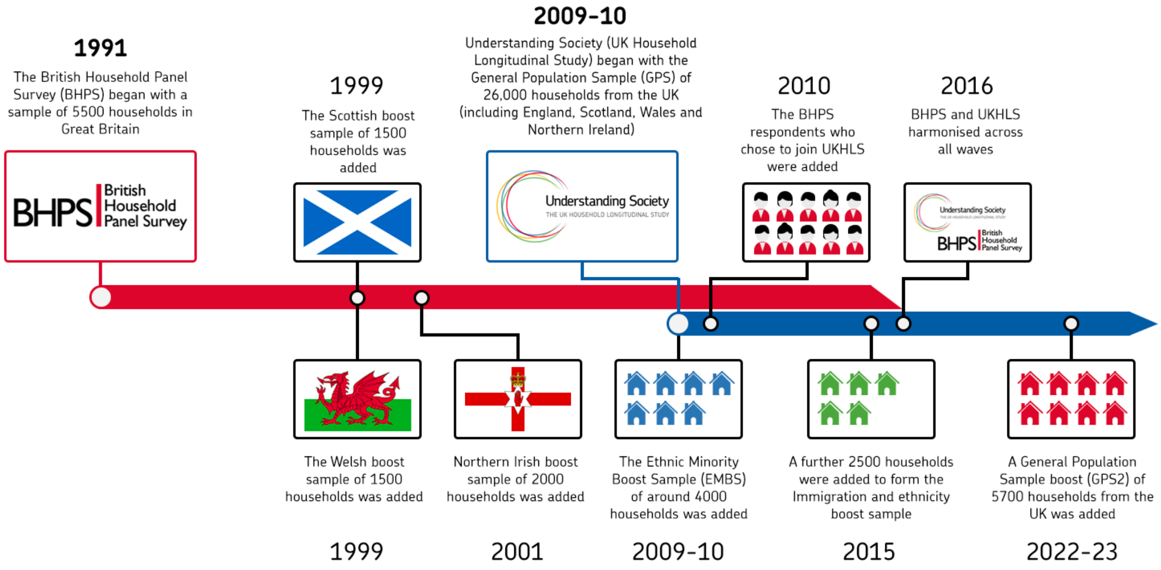
While the samples are probability samples, not all sections of the population were selected with the same probability. Also, not everyone selected and asked to participate in the interviews did so. To correct for bias due to these two reasons we recommend you use the weights provided. **TIP:** This is explored in the weighting and sampling section.

All samples other than the GPS-NI and GPS2-NI part had a clustered and stratified design. As statistical softwares assume that the data is a simple random sample, to estimate standard errors correctly you will need to explicitly inform the software about the clustering variable **w_psu** (primary sampling unit) and the stratification variable **w_strata**. For guidance on this visit the Clustering and stratification page.

Survey timeline

Understanding Society builds on its predecessor project, the British Household Panel Survey (BHPS).

Many design features, instruments, and questions from the BHPS were continued in Understanding Society allowing analysts to combine data from the two datasets. The active BHPS sample were incorporated into Understanding Society from Wave 2 onwards. Since 2017, a harmonised version of the two surveys has been released to make this easier.


1991
The British Household Panel Survey (BHPS) began with a sample of 5500 households in Great Britain
BHPS British Household Panel Survey
1999
The Scottish boost sample of 1500 households was added
2009-10
Understanding Society (UK Household Longitudinal Study) began with the General Population Sample (GPS) of 26,000 households from the UK (including England, Scotland, Wales and Northern Ireland)
Understanding Society THE UK HOUSEHOLD LONGITUDINAL STUDY
2010
The BHPS respondents who chose to join UKHLS were added
2016
BHPS and UKHLS harmonised across all waves
The Welsh boost sample of 1500 households was added
1999
Northern Irish boost sample of 2000 households was added
2001
The Ethnic Minority Boost Sample (EMBS) of around 4000 households was added
2009-10
A further 2500 households were added to form the Immigration and ethnicity boost sample
2015
A General Population Sample boost (GPS2) of 5700 households from the UK was added
2022-23

Sample members are interviewed every year as long as they continue to live in the UK and can be located, contacted and agree to participate. The sample is issued in monthly batches across 24 months for the GPS-GB, GPS2-GB and EMBS but across the first 12 months for GPS-NI, GPS2-NI and BHPS, and across the last 12 months for IEMBS. From 1991-2009, the BHPS fieldwork period was from September to December of each year. View the Survey's timeline and information about Data releases.

Although GPS-GB and EMBS samples are in the field for over 24 months every individual is interviewed at approximately 12 month intervals. So, for any individual their Wave 2 interview will be approximately one year after their Wave 1 interview, and their Wave 3 interview will be around 1 year after their Wave 2 interview (and two years after their Wave 1 interview) and so on. These two aspects result in overlapping fieldwork periods, i.e., in any year interviews for two consecutive waves are conducted. In 2010, part of the GPS and EMBS samples were interviewed for the first time (Wave 1, Year 2 sample) and part was interviewed for the second time (Wave 2, Year 1 sample), and so on. The same will apply for the GPS2-GB sample that was selected and interviewed for the first time in the 24 month period, 2022-23.

Tips for analysts:

To compare and analyse differences across two consecutive calendar years, you should use data from all the interviews conducted in those years across two waves. For example, if you want to compare 2010 and 2011, you will need to use data from Wave 1 and Wave 2 interviews that were conducted in 2010 and compare that data with the data from Wave 2 and Wave 3 interviews that were conducted in 2011.

Analysing changes during the pandemic; comparisons of calendar year data

The analysis guidance described above also apply to researchers interested in analysing changes during the pandemic (2020) compared to the pre-pandemic period (2019). To make it easier for researchers and analysts to conduct this type of analysis we have released a 2020 calendar year dataset (which includes interviews conducted in 2020 as part of Wave 11 and Wave 12). New cross-sectional calendar year datasets are planned for each subsequent year and are explained in the Data section under comparisons of calendar year data. These are available under a separate study number and meant to be used for cross-sectional analyses and trends.

Note that while you can produce the 2019 calendar year dataset by combining interviews conducted in 2019 as part of Wave 10 & 11, this dataset is available with the Covid-19 survey data, the new monthly (then bi-monthly) survey that was fielded to the main survey sample members from April 2020 to September 2021 to understand the impact of the pandemic on people's lives.

Interview process

Interviews are typically carried out face-to-face in respondents' homes by trained interviewers or by respondents themselves completing their survey online. Every question is answered voluntarily.

All household members of the households selected at the first wave and their descendants constitute the core sample and are followed wherever they move within the UK to see how things have changed over time and over their lifecourse. All those who join their households in subsequent waves do not become part of the core sample but are interviewed as long as they live with at least one core sample member. For more technical details on these see the Following rules section below.

The **household enumeration grid** identifies household members and collects some basic information about them. Any knowledgeable adult in the household can fill this in.

The **household questionnaire** is generally asked of the person who owns or rents the accommodation. This questionnaire includes questions relevant to the whole household such as expenditure on heating, information about ownership, mortgage, rents and so on.

All household members aged 16 or above, are asked to complete an **individual questionnaire**. During face-to-face interviews, interviewers ask most of the questions, but respondents complete one section, the self-completion questionnaire (this was available as a paper questionnaire for the first two waves after which it was available on the computer (CASI)).

Household members aged 10-15 years are asked to complete a short **self-completion youth questionnaire**, with permission from their parent or carer. They become eligible for a full interview once they reach the age of 16.

Some information about children **0-9 years old** is collected from their parents or guardians.

Interviewers also provide some information about the condition of the property, cooperativeness of the respondents and any difficulties faced by respondents in answering the questions etc.

Who answers which questions?

This video describes which respondents answer which questions and questionnaires and by which mode during their interviews for Understanding Society.

Mode – Web, CAPI and telephone

In the BHPS (from Wave 3) and for the first six waves of Understanding Society*,* interviews were mostly conducted face-to-face. From Wave 3 in the BHPS and from Wave 3 of Understanding Society onward, a small number of respondents are interviewed over the phone. From Wave 7 onwards, web interviewing was introduced. At Wave 7, it was only adults in households that did not take part in Wave 6 that were invited to complete online. From Wave 8, adults in some households that had taken part in Wave 7 were also invited to complete online. While 20% of the sample was ring-fenced to be interviewed in person, 40% were invited to complete their survey online, and the remaining 40% of the sample were issued directly to interviewers. In each successive wave (until 2020) the proportion of the sample invited to take part online increased to a maximum of 70%. Adults who are invited to complete online, but do not take part in the first five weeks, are issued to interviewers who then try to contact them to take part in-person. Adults issued directly to interviewers who do not take part in the first ten weeks are invited to complete online. Both web-first and CAPI-first non-responders were also eligible for the telephone “mop-up” towards the end of the fieldwork period. This mode is used to increase participation among those adults who are hard to contact in person.

In mid-March 2020, face-to-face interviewing was suspended due to the ‘lock-down’ associated with the COVID-19 pandemic. Adult sample members who were allocated to interviewers at that time were sent a letter, explaining the position with face-to-face interviewing, and giving them their log-in details so they could complete online. Those who did not complete online were followed up by interviewers who tried to conduct the interview by telephone. From the April 2020 monthly sample onwards, all adult sample members were issued web-first, with telephone as the follow-up mode. As interview mode is known to have an impact on responses, users are advised to read the section Analysis advice for mixed mode data. Face-to face interviewing was resumed from April 2022.

Tips for analysts: COVID-19

As noted above, with the arrival of COVID-19, all face-to-face interviews were suspended and we invited all our sample members to take part online or by telephone. Some new questions were added about COVID-19 and furlough. Face-to-face interviews were used again from April 2022. We have brought together a document to help researchers explore Understanding Society changes to the main study due to the COVID-19 pandemic.

Interview dates, fieldwork periods

The sample is issued monthly. The start date and length of the fieldwork period has varied across waves (earlier waves having shorter fieldwork periods), interview modes and samples. The fieldwork period is generally split into an initial interviewing phase, and then a shorter ‘re-issue’ phase where non-responding households are reviewed and re-issued, often to a different interviewer. To find out more about fieldwork start dates and durations see Understanding Society fieldwork procedures.

The Great Britain part of the GPS, GPS2 and the EMBS comprise of 24 monthly samples while the Northern Ireland part of the GPS, GPS2 and the BHPS samples comprise of 12 monthly samples. The IEMB sample, which was added in Wave 6, was issued as 4 quarterly samples. For the GPS-GB, GPS2-GB part and EMBS the fieldwork is around 24 months, while for the BHPS, GPS-NI and GPS2-NI part it is during the first 12 months of a wave and for the IEMBS it is during the last 12 months of a wave. While the sample month may change for sample members across waves, the sample quarter remains the same across waves.

Note that even though the fieldwork runs over two years for some samples, and the duration of fieldwork for sample months change across waves and modes, every sample member is interviewed approximately one year apart. View the Survey’s timeline.

Information about the sample month in which a household was issued is available in the variable **w_month**, which is located in a number of data-files: **w_hhsamp**, **w_indsamp**, **w_issue**, **w_callrec**, **w_child**, **w_hhresp**, **w_indresp**, and **w_indall**, **xwaveid**. The variable that identifies the sample quarter is **w_quarter** in these files and **xwavedat**. The interviews for BHPS Waves 1-18 were conducted generally between September and December of each year with the re-issue period extending for up to 5 months after that. The BHPS sample was issued as a single batch, so this information is not applicable and hence not available for the BHPS Waves 1-18.

Due to the long fieldwork period for each monthly sample, the month issued will not necessarily be the month in which the interview took place. The actual date of interview for the individual adult interview is given in the **w_indresp** file in variables **w_istrtdatd** (day), **w_istrtdatm** (month), and **w_istrtdaty** (year). The variables for the date of the household interview are **w_histrtdatd** (day), **w_histrtdatm** (month), and **w_histrtdaty** (year).

Information about the issue number, whether first issue or a subsequent re-issue, is available in the **w_issueno** variable, located in **w_callrec** data file (all waves), and the **w_issue** data file (Waves 1-5 only). These files also include information about all the calls made until a successful interview and the outcomes of each call. This information is not available for the BHPS Waves 1-18.

Incentives

All incentives are per person and are a token of our appreciation for their help. Most of the incentives are unconditional. Over the years to improve response rates, evidence-based changes have been made to the incentive schemes. The current design is that for every adult household member who participated in the last wave and those who turned 16 in the current wave, the invitation letter to complete the survey includes an unconditional £20 gift voucher. Those respondents who turn 16 are welcomed to the adult survey with an additional incentive of being entered into a prize draw to win an iPad. All 10-15-year-olds are sent or given an unconditional £10 incentive with their paper self-completion questionnaire. These are either handed over by the interviewer or sent through the post, depending on the mode in which the household grid is completed. Adult sample members who did not take part during the previous wave, including new adult entrants to the household, are offered a £20 incentive if they are able to take part in the current wave (conditional). For adults in web-first households there is a bonus £10 incentive which is sent to them if they complete the survey within five weeks (conditional).

Consent

Every question, in each section of the questionnaire, is answered voluntarily by participants. Information on the consent process is outlined on the consent information page. Understanding Society asks for consent to link responses to certain records held by government departments and other agencies. The type of records includes health, education, benefits records etc. and vary according to the wave and whether consent has already been sought either verbally or in writing. Those asked for consent are given a copy of the consent information leaflet along with a privacy notice to read. Read more about linked data.

Fieldwork Documents

The fieldwork documents used during the interviewing process include emails, letters, leaflets and project instructions for interviewers. Details of who answered which module is listed in the Modules table within the "Project instructions for Interviewers" documents (Waves 1-8), there is a separate one for each wave.

Following rules

Everyone enumerated in the households when a sample is selected are considered to be an Original Sample Member (OSM). Anyone who moves in with an OSM from onwards the subsequent wave is considered to be a Temporary Sample Member (TSM). The only exception is any non-ethnic minority individual enumerated in EMBS households in Wave 1 and in IEMBS households in Wave 6. They are considered to be TSMs. Any child born to an OSM mother is an OSM. Any TSM who has a child with an OSM becomes a Permanent Sample Member (PSM). Note, in the BHPS children of OSM mothers and OSM fathers became OSMs. But the new rules are applied to any children born to BHPS sample members after they became part of Understanding Society.

OSMs and PSMs are always eligible for interviews as long as they are living in the UK. TSMs are eligible for interviews only if they live with at least one OSM or PSM. This means that when issuing the sample at the next wave, households which had been identified as containing only TSMs at the previous wave are not issued.

There are other reasons that a household may not be issued at a particular wave, depending on information received between waves. Households are withdrawn if we are informed that a whole household has adamantly refused or asked to withdraw from the Study; have emigrated; have died; or are no longer mentally or physically capable to make an informed choice to consent.

From Wave 4 onwards, until the COVID-19 pandemic, we also withdraw households as ‘dormant’ for which the outcome at the previous two consecutive waves were both non-contact, were both refusal, or there was a refusal two waves prior followed by a non-contact one wave prior. The rules to designate households as dormant were suspended after the COVID-19 pandemic. During this time, it had not been possible to interview sample members face-to-face, and so keeping the dormant rules may have resulted in people who were not able or willing to take part online or on the phone to have been dropped from the sample.

Fieldwork procedures (2009 -)

Information about the Understanding Society fieldwork procedures: the fieldwork period and how it has changed across waves and differs by interviewer mode, panel maintenance rules.

Fieldwork dates

For the first five waves of the Study, each monthly fieldwork sample started on the 8th of the month, except in Northern Ireland where it started on the 1st. August and December months started a week earlier, on the 1st, to allow interviewers more time to make contact during the holiday periods. At Wave 6, each monthly fieldwork period started on the 1st, except for January which started on the 8th, to avoid fieldwork agency staff working during the Christmas period. At Wave 7, to make it easier for the fieldwork agency to deal with the logistics of issuing two waves, the start date for each month was moved to the 8th. From then, the start date for each Wave alternated between the 1st and 8th of the month.

With the wide introduction of web as a primary mode of interview, the start date for each sample month was moved back to before the start of the month. This was to allow any non-responding sample members who had been invited to complete online to be issued to interviewers at the same time as the CAPI-first sample. This avoided the situation of interviewers starting to work their caseload for a few weeks before then being given the web-first non-responding cases to work. Having their whole caseload at the start of the CAPI fieldwork meant that interviewers could manage their work more efficiently.

In the first year of Wave 8, the web-first fieldwork started three weeks before the CAPI fieldwork, and by the start of the second year this had increased to five weeks before the CAPI fieldwork. The web fieldwork started one week before the start of the month, which meant that the CAPI fieldwork started around the start of the next month. For example, the February 2017 sample at Wave 8 were invited to take part online on January 25th, and the CAPI fieldwork started on March 1st.

Fieldwork periods

At Wave 1 the initial fieldwork period for first issue was one month, starting on the 8th of each month. Then there was a one week pause when the non-completed sample was reviewed, and then a two week re-issue period. For the EMB, the screening period took place two weekends before the start of the first issue and covered six days (Wednesday to Monday). No interviewing could be done if a household was screened-in until the 8th of the sample month. During Wave 1 there were a number of changes to the fieldwork periods. For the EMB the screening period was doubled from six to twelve days from July 2009 (year 1) to try to reduce the high non-contact rate during the screening process. At the start of year 2 (January 2010), the first issue fieldwork period was extended from four to six weeks to allow interviewers more time to cover their assignment and reduce non-contact rates.

The second wave of Understanding Society was the first longitudinal wave, and so interviewers had to deal with sample members moving house and tracing them to a new address. Therefore, to allow time for this new task there was an additional four week period to enable interviewers to trace movers, either through their own efforts or where a new address had been identified by in-office tracing and reissued to the interviewer. From April of year 1 (2AP1), the reissue fieldwork extended to four weeks, from two. This meant that the fieldwork period was now four months long.

At Wave 4 the fieldwork period changed from that used at Waves 2 and 3. In the first year of Wave 4 the initial six weeks of main issue, followed by a two week gap for office administration, was followed by a two week CAPI reissue phase and then a six week CAPI mop-up to allow extra time for tracing movers, and then there was a four week CATI mop-up, where the first week overlapped the sixth week of the CAPI mop-up. This resulted in a 19-week fieldwork period. In year 2 of Wave 4 this was adapted to lengthen the issue periods. The initial main issue was increased to ten weeks with an additional two week extension to cover appointments and traced movers. The in-office administration and tracing gap was removed, but this work was done on an ongoing basis and cases were reissued in two batches, at weeks eight and 10. The CAPI reissue and CAPI mop-up stages went on until week 16, with an additional three week period for appointments. The last four weeks of this period ran in parallel to the CATI mop-up (weeks 16-19).

At Wave 6 there were more significant changes to the fieldwork design due to the change in fieldwork agencies following the competitive procurement process for Waves 6-8. These were implemented because it was thought that they would improve response on the Study. The fieldwork period was increased from 19 to 23 weeks; an initial issue of eight weeks with a two week mop-up period, then three weeks whilst sample was returned to the operations department and reviewed for reissues, a six week CAPI reissue period, then a four week CATI mop-up period, although during these last four weeks some face-to-face interviewing could still be done. There was some flexibility around these timings so interviewers could retain households after 10 weeks if they had a strong chance of interviewing them, such as where there are appointments.

At Wave 7 the 23-week fieldwork period for the web-first households was broken up into two weeks of 'web-only' fieldwork, 17 weeks of CAPI fieldwork (with the web still available), then four weeks of the telephone mop-up. Adults in households that had responded at Wave 6 were issued directly to interviewers, with a 19-week CAPI fieldwork period followed by a 4-week telephone mop-up. From Wave 8, the web-only fieldwork period was extended to five weeks. During the re-issue phase for these households, non-responding adults were sent a re-issue letter which included information on how to complete their interview online (week 15 onwards).

During Wave 11 fieldwork, the lockdown associated with the COVID-19 pandemic meant that face-to-face interviewing was suspended (mid-March 2020). From that date, all sample members were issued web-first, with those who had not responded in the first five weeks of fieldwork allocated to interviewers who tried to contact and interview them by telephone (CATI). Thus, there was no longer a final telephone mop-up period. For further details please see the document "Understanding Society changes to the main study due to the COVID-19 pandemic".

Panel membership and panel maintenance

Prior to fieldwork starting, ISER transfers the sample information required to the fieldwork agency. The sample files contain information on the individuals and households being issued to field, along with any prior wave information used in dependent interviewing, such as previous occupation. From Wave 4 onwards, the sample fed-forward for each sample month also includes information on 'dormant' households. These are households where the household outcome at the previous two consecutive waves were non-contact, refusal, or a refusal (T-2) followed by a non-contact (T-1). These households are considered to be no longer active in the survey. They are issued to the fieldwork agency, but not to interviewers. This is in case a household member of a dormant household contacts ISER and requests to be part of the Study again.

Response rates

Calculation for response tables

The response rates reported in these tables are designed to measure the effectiveness of fieldwork, and so take as the base those cases that were issued at each wave. This means that cases which were withdrawn before fieldwork started (e.g., if a household contacts us to withdraw from the study in between waves) are not included in the tables.

Effects of mode transition on response rates

In summary, to establish the effects of mode transition during the pandemic, comparisons were made between the 2019 and 2020 samples for the period April to December. This found around three-quarters of those who had completed in CAPI in 2019 took part in 2020 using a different mode. Around one-quarter of those who had not responded in 2019, did respond in 2020. Overall, the response rate for the 2020 sample was just 1.1 percentage points lower than the response in 2019. Response in 2020 was lower among those in the low web propensity sample. Response in 2020 was particularly lower for those in the higher age groups, those who live alone, and those with lower levels of education. An analysis of the unweighted sample composition indicated that there were significant differences in the responding sample in 2020, but that these differences were relatively small, with most under 2 percentage point differences. However, researchers should be aware of the potential for these differences to affect analyses and so use the correct weights or control for factors which may affect response in their models. Further details of these changes are included in the document mentioned below and detailed analyses of mode transition and response behaviour is available in (Alvarez et al., (2021).

Tips for analysts: COVID-19

We have brought together a document to help researchers explore Understanding Society changes to the main study due to the COVID-19 pandemic.

Household response

The **w_hhsamp** file is used to create the household outcome summary variable. The household interview outcome variable (**w_ivfho**) code 10 (all eligible hh intv) is counted as “Fully responding households” while codes 11 (interviews + proxies), 12 (interviews + refusal), 14 (hh grid + indiv only (no hh ques)) are counted as “partially responding households” (i.e., at least one adult interview). Codes 51-59 (no hh member contact, unable to locate address, contact made but not with correct, address inaccessible, no phone contact with CATI hhold, unknown eligibility, other non-contact) are counted as “Non-contact” households. “Untraced movers” are households with households response outcome code 50 (address not found). Codes 60 (refuse to rsrch cntre), 61 (refusal to intviewer), and 66 (no web response/web-only hh) are counted as “Refusal” households. The “Other non-interview households” category is made up of codes 15 (hh grid + proxies only), 16 (hh grid only), 62 (language problems), 63 (no intv.: age/health), and 65 (ill during survey period). Households with outcome codes 70-98 are counted as ineligible and not included in the tables.

Individual response

The **w_indsamp** file is used to create the individual outcome summary variable. The individual interview outcome variable (**w_ivfio**) code 1 (full interview) is counted as “full interview”, and code 2 (proxy interview) as “proxy interview”. The “other non-interview” category includes sample members with individual interview outcome codes 9 (lost capi interview), 11 (other non-intvw), 14 (ill/away during survey period), 15 (too infirm/elderly), 16 (language difficulties), and 18 (unknown eligibility). Code 10 (refusal) is used for the refusal category. All the outcome codes that refer to children are coded as ineligible for the adult response tables. Also ineligible are codes 80 (tsm - no osm/psm), 81 (prev wave adamant refusal), 82 (l-t untrace, w-drawn), 83 (withdrawn before field), 84 (other ineligible), 98 (other retiring (due to health)), and 99 (dead). Other outcome codes that refer to specific outcomes within a non-interviewed household are replaced with their household response variable (see above) and that is used to categorise them as refusals, non-contacts or other non-interview (e.g., the individual outcome is coded as a household non-contact if the household outcome is non-contact).

Data

This video gives more detail on how Understanding Society data is structured and gives basic information about how the data is collected.

Download the EUL version of Understanding Society data direct from the UK Data Service. Visit our Data Releases page to see the latest data available and when further updates are expected.

Variable naming convention

Understanding Society has a distinct naming convention for its data files to identify which wave the data is from and the source of the data.

The naming convention for variables follows the same rules as file names. Variable names have the same root name which is fixed over time, and begin with a prefix to reflect the wave the data are collected (“**a_**” for the first wave, “**b_**” for the second wave; in this user guide we have used “***w_***” to denote waves in general). For example, current employment status collected from interviews with responding adults in Wave 1 (both years: 2009 and 2010) is **a_jbstat** and **b_jbstat** in Wave 2 (both years: 2010 and 2011).

To ease identification of groups of variables a number of additional general naming conventions have been applied. For instance, following the wave prefix, information from the UKHLS Wave 1 and Wave 2 self-completion interview with adults starts with the prefix “**sc**”; information from the interview with young adults generally starts with the prefix “**ya**”, and information from the child development module starts with the prefix “**cd**”. Similarly, we have attempted to include in the variable name the acronym of well-known instruments such as the Strengths and Difficulties Questionnaire (SDQ) or the General Health Questionnaire (GHQ). See, for example, **c_ypsqda** to **c_ypsdqy** on data file **c_youth** or **d_scghq2_dv** on data file **d_indresp**.

The prefix “**ff_**” following the wave prefix shows variables that were fed-forward from previous waves to route respondents appropriately in the questionnaire. To aid data collection, information reported at an earlier time is fed-forward to the respondent to personalize the question. Rather than ask a question about changes in marital status (mstatsam), the question might say:

“When we last interviewed you, your legal marital status was [ff_marstat] . Has your legal marital status changed at all since [ff_intdate] ?”Feed-forward variables are used at both the household and individual levels. For example:

b_ff_hhsize feeds forward the household size from the previous wave (Wave 1)

b_ff_plbornc is the country of birth of the respondent fed-forward from the previous wave.

Note that the variable name does not change over time so long as the underlying question does not change substantially. Analysts are advised to carefully read the variable notes in the online documentation to keep track of any definitional changes or changes in the code frame that may impact study results. An example is the derived variable **w_qfhigh_dv** which provides limited information about continuing BHPS (from Wave 2 onward) and IEMB sample members (from Wave 6 onward) as the underlying code frames for the initial conditions questions in the BHPS Wave 1-18, UKHLS Wave 1-7 and IEMB Wave 1 (as part of UKHLS Wave 6) do not perfectly align.

Missing values

For various reasons survey responses may not have a valid code or value. The missing value codes assigned to these data are described below. All missing values are negative and are never used as valid responses. We recommend that users carefully read the questionnaires and compare missing value distributions across waves before using the substantive information contained in them.

Value	Description
-1	"Don't know" - When the respondent does not know the answer.
-2	"Refused" - When the respondent does not know the answer.
-7	"Proxy" - Sometimes when a person cannot participate in the interview, someone else in the household (generally their spouse or partner or adult children) answers questions on their behalf, that is, by proxy. This questionnaire is a much shorter questionnaire asking factual information. So, if a question was not included in the proxy questionnaire and the person gave a proxy interview, this variable will be missing for them. In such cases the variable will have a value of -7.
-8	"Valid skip" - This information is missing because the person was never asked this question as they were not eligible for it. E.g., someone who is not in paid employment is not asked questions about their pay.
-9	"Missing by error or implausible".
-10	"Not available for the IEMBS" - Some questions were only asked in the W6 questionnaire for non-IEMBS and so individuals in IEMBS will have missing information for these variables.
-11	"Only available for the IEMBS" - Some questions were only asked in the W6 questionnaire for IEMBS and so individuals in non-IEMB samples will have missing information for these variables.
-20	"No data from the BHPS W1-18" - This code is only used for variables in the xwavedat file which is harmonised across BHPS and UKHLS. If some variable was only asked in the UKHLS then there will be no data from BHPS W1-18 and hence missing for those not interviewed during UKHLS.
-21	"No data from the UKHLS" - This code is only used for variables in the xwavedat file which is harmonised across BHPS and UKHLS. If a variable was only asked in the BHPS then there will be no data from UKHLS and hence missing for those not interviewed during BHPS W1-18.

Note that the default missing value code for post-field derived variables tends to be "missing or wild". This also applies to most variables on the **xwavedat** file. Missing value codes on the youth self-completion questionnaire also tend to be less accurate because the instrument was administered as a paper-and-pencil questionnaire and so it is not clear whether they refused to answer, didn't know the answer or simply missed the question. They may also not have followed the question routing correctly.

Tips for analysts

Income variables can be negative due to self-employment reported losses. A value of 0 means their income was 0 and negative values (other than those dedicated for missing values) are actual negative values reported. All individuals with negative net labour income are self-employed. In this survey data, missing values are assigned a negative value. These negative values are dedicated for the missing values and will never reflect actual negative income values reported. **Note:** users can identify those self-employed using the variable **w_jbsemp.**

List of data files and their descriptions

All data files released under the main study SN6614 EUL version are listed below. The Special License version of the survey SN6931 includes the same files but with some additional variables and for some income variables the non-top-coded values of those variables. The Secure Access version of the survey data is available as SN6676. It includes all files in the Special Licence version and files that contain 3 variables relating to the National Grid Reference for each household: Easting, Northing and positional quality indicator (OSGRDIND). This Secure Access version also includes variables for the full dates of birth for Understanding Society and BHPS respondents. The different access levels are explained on the Data Access page.

Data collected from different sources (e.g., the household interview, the adult interview, the youth interview) are stored in separate files. Each wave has a set of such files. To make it easier to use, files have the same root name, but begin with a letter prefix to reflect the wave the data are collected. So, "**a_**" for the first wave, "**b_**" for the second wave (in this user guide we have used "***w_***" to denote waves in general). From the Wave 7 data release onward (November 2017) Understanding Society-harmonised BHPS data files are also included. Most files exist for both studies and if they do and have been harmonised, the file stem name will match. Wave-specific harmonised BHPS files can be identified by the wave prefix ***bw_***. Note: BHPS data files that have not yet been harmonised but have the same stem name as UKHLS files have the suffix ***_bh*** , but files that are unique to the BHPS do not need to have such suffixes (harmonised or not).

Table 1 lists the main data files such as **w_indresp** which includes information collected during adult interviews, **youth** which includes information collected during youth interviews. To avoid creating very large files, some information collected during adult interviews are provided as separate smaller multi-level levels, see **Table 3** and **Table 4** for a list of these files.

Stable characteristics – information collected once during the first interview

Some stable information such as date of birth, ethnicity, country of birth is collected in the first time a person is interviewed. So, while for the core sample members this will be asked in the initial wave, for those joining the household after that, they will be asked in the wave they joined. To make it easier for data users, this information has been asked in different waves for different respondents and has been put in one individual level file, **xwavedat**. There are a few other such cross-wave files, which all begin with "**x**". See **Table 2** for a list of such files.

Some information is collected about the interview and sampling process, such as number of calls made by the interviewer, outcome of each call, interviewer ID, the information the interviewer collects about the condition of the property and neighbourhood, time taken to complete a questionnaire module and so on. See **Table 5** for a list of these files.

Some datafiles are particularly useful for analysts, irrespective of their area of research.

Table 1: List of main data files

Filename	**Description**
w_indall bw_indall	Household grid data for all persons in household, including children and non-respondents. *The variable pidp or the combination of variables "w_hidp w_pno" uniquely identifies each row of w_indall. The variable pidp or the combination of variables "bw_hidp bw_pno" uniquely identifies each row of bw_indall.*
w_hhresp bw_hhresp	Substantive data collected from responding households. *The variable w_hidp uniquely identifies each row in b_hhresp. The variable bw_hidp uniquely identifies each row in bw_hhresp.*
w_indresp bw_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). *The variable pidp or the combination of variables "w_hidp w_pno" uniquely identifies each row of w_indresp. The variable pidp or the combination of variables "bw_hidp bw_pno" uniquely identifies each row of bw_indresp.*
w_youth bw_youth	Substantive data from youth questionnaire. *The variable pidp uniquely identifies each row in these files. The variable pidp or the combination of variables "w_hidp w_pno" uniquely identifies each row of w_youthl. The variable pidp or the combination of variables "bw_hidp bw_pno" uniquely identifies each row of bw_youth.*
w_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. *The variable pidp or the combination of variables "w_hidp w_pno" uniquely identifies each row of w_child.*
w_egoalt bw_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. *The combination of variables "pidp apidp" or "w_hidp w_pno w_apno" uniquely identifies each row in w_egoalt. The combination of variables "pidp apidp" or "bw_hidp bw_pno bw_apno" uniquely identifies each row in bw_egoalt.*
w_income bw_income	This file contains reports of unearned income and state benefits for each individual. *The combination of variables "pidp w_fiseq" uniquely identifies each row in w_income. The combination of variables "pidp bw_ficode bw_fiseq" uniquely identifies each row in bw_income.*

Table 2: List of cross-wave files

Filename	Description
xwavedat	Stable characteristics of individuals, such as date of birth, country of birth, ethnicity, which is typically collected only once in the lifetime of the Study are picked from different data files and put into this file. This file now includes all sample members ever enumerated in either Understanding Society and BHPS and variables have been harmonised across studies where possible. *The variable pidp uniquely identifies each row.*
xivdata xivdata_bh	Some basic information about interviewers is stored in these files. [non-harmonised] These are available in the Special License version of the survey SN8579.
xwaveid xwaveid_bh	Some basic sampling information from each wave such as interview outcomes is included in this file. [non-harmonised] *The variable pidp uniquely identifies each row in these files.*
xwlsten	Contains information on the latest known sample status of individuals [Only BHPS]. *The variable pidp uniquely identifies each row.*
xhhrel	Family matrix file which allows family members and households to be connected over time. The file compiles existing Understanding Society main survey data, particularly from the egoalt and indall files, and includes every sample member ever enumerated as part of the study. *The variable pidp uniquely identifies each row and the variable osm_hh identifies the cross-wave household each pidp belongs to.*

Table 3: List of data files about children based on information collected during adult interviews

Filename	Description
a_natchild f_natchild n_natchild2 bb_childnt bk_childnt bl_childnt	Some basic information about all **biological children** born to the sample members, whether co-resident or not. These are collected in the first wave for any sample. So, for example, **a_natchild** was collected in Wave 1 of UKHLS for GPS, EMBS and **bl_childnt** was collected in Wave 12 of the BHPS for the Northern Irish boost sample. These files are not harmonized. *The combination of variables "pidp w_childno" or "w_hidp w_pno w_childno" uniquely identifies each row of the files w_natchild. The combination of variables "pidp bw_lncno" or "bw_hidp bw_pno bw_lncno" uniquely identifies each row of the files bw_childnt.*
a_adopt f_adopt n_adopt n_stepchild	Some basic information about all **adopted and stepchildren** born to the sample members, whether co-resident or not. These are collected in the first wave for any sample. So, for example, **a_adopt** was collected in Wave 1 for GPS, EMBS and **bl_childat** was collected in Wave 12 of the BHPS for the

bb_childad bk_childad bl_childad	Northern Irish boost sample. These files are not harmonized. Note that in Wave 14, information about stepchildren for GPS2 was collected separately from that of adopted children and is available in a separate file **n_stepchild**. *The combination of variables "pidp w_adoptno" or "w_hidp w_pno w_adoptno" uniquely identifies each row of the files w_adopt. The combination of variables "pidp bw_lacno" or "bw_hidp bw_pno bw_lacno" uniquely identifies each row of the files bw_childad.*
w_newborn	Every wave after Wave 1, basic information about **new born children** such as brithweight, etc. is collected from new parents. *The combination of variables "pidp w_newchno" or "w_hidp w_pno w_newchno" uniquely identifies each row of the files w_newborn*
w_chmain	Information about **child maintenance** arrangements was collected in Waves 3, 5, 7, 9, 11,…. *The combination of variables "pidp c_absparno" or "c_hidp c_pno c_absparno" uniquely identifies each row in c_chmain. The combination of variables "pidp w_childpno" or "w_hidp w_pno w_childpno" uniquely identifies each row in w_chmain where w is e, g, I, k,..*
w_parstyle	Every wave from onwards Wave 4, information about **parenting styles** was collected. *The combination of variables "pidp w_childpno" or "w_hidp w_pno w_childpno" uniquely identifies each row of the files w_parstyle*

Table 4: List of data files about partnerships, jobs and employment histories based on information collected during adult interviews

Filename	Description
a_marriage f_marriage ba_marriag bk_marriag bl_marriag	Start and end dates of past marriages and how that marriage ended was collected during adult interviews in the first wave a sample was selected. So, for example, **a_marriage** was collected in Wave 1 for GPS & EMBS. [non-harmonised] *The combination of variables "pidp w_marno" or "w_hidp w_pno w_marno" uniquely identifies each row of the files w_marriage. The combination of variables "pidp bw_marno" or "bw_hidp bw_pno w_bmarno" uniquely identifies each row of the files bw_marriag.*
a_cohab f_cohab bb_cohabit bk_cohabit bl_cohabit	Start and end dates of past cohabitations and how that cohabitation ended was collected during adult interviews in the first wave a sample was selected. So, for example, **a_cohabit** was collected in Wave 1 for GPS & EMBS. [non-harmonised]. *The combination of variables "pidp w_cohabno" or "w_hidp w_pno w_cohabno" uniquely identifies each row of the files w_cohab. The combination of variables "pidp bw_lcsno" or "bw_hidp bw_pno bw_lcsno" uniquely identifies each row of the files bw_cohabit.*
bw_jobhist (bw_jobhistd)	Contains information about employment history between two waves collected during adult interviews. [Only BHPS]. *The combination of variables "pidp bw_jspno" or "bw_hidp bw_pno bw_jspno" uniquely identifies each row in these files.*

a_empstat e_empstat bw_lifemst bk_lifemst bl_lifemst	**Employment history** was collected during adult interviews in Wave 1 for part of the GPS & EMB samples and in Wave 5 for rest of the samples, this was not asked for the IEMBS in Wave 6. [non-harmonised]. *The combination of variables "pidp w_spellno" or "w_hidp w_pno w_empstat" uniquely identifies each row of w_empstat. The combination of variables "pidp bw_leshno" or "bw_hidp bw_pno bw_leshno" uniquely identifies each row of bw_lifemst.*
bc_lifejob	Contains information about jobs held in employment spells [Only BHPS]. *The combination of variables "pidp bw_ljseq" or "bw_hidp bw_pno bw_ljseq" uniquely identifies each row of this file.*

Table 5: Paradata and interview related files

Filename	**Description**
w_hhsamp bw_hhsamp	This file contains information about each household that the interviewer collects about the condition of the property, neighbourhood, interview outcome and so on. *The variable w_hidp uniquely identifies each row in b_hhsamp. The variable bw_hidp uniquely identifies each row in bw_hhsamp.*
w_indsamp bw_indsamp	Includes current interview outcome for anyone enumerated in the last interview wave, for example, whether they have responded, only enumerated, couldn't be contacted or refused, or were ineligible. *If you restrict the data to cases where w_finloc=1/bw_finloc=1, then pidp uniquely identifies each row.*
w_callrec	Includes information about each interview call made to each household, such as outcome of the call, interview ID. *The combination of variables "w_hidp w_issueno w_callno" uniquely identifies each row in this file.*
Timings Files	Various files are available that capture the time taken to complete questions and modules within individual and household questionnaires. Given that these files vary from wave to wave and are of limited, specialist use only, they are not released as standard. If you want to make use of them please contact usersuport@understandingsociety.ac.uk who will be happy to advise.

Linking data files

Understanding Society covers the whole household and researchers may want to link household members within waves and across waves.

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, **a_hhresp** can be linked with **a_indresp** using **a_hidp** but not with **b_hhresp**.

Individuals are identified by the personal unique identifier (**pidp**), which is the same in all waves and can be used to link information about a person from different records belonging to one wave, or to link information from different waves. The cross-wave person identifier **pidp** is consistent across the harmonised BHPS and UKHLS files. Additionally, individuals are identified by ***w*_pno** – the person number within the household. The combination of ***w*_hidp** and ***w*_pno** is unique for each individual only in data files for wave w. For example, **ba_indresp** can be linked with **h_indresp** using **pidp**; **a_indresp** can be linked with **a_indall** using **pidp** or using **a_hidp a_pno**. But **a_indresp** cannot be linked with **b_indresp** using **a_hidp a_pno**.

To help researchers link families we have created a family matrix data file (**xhhrel)**. This data file identifies all family relationships across the Study, allowing researchers to see where people and households connect. For further information read the Connecting family members and households over time web page.

Tips for analysts:

Take a look at the syntax for help in linking household members and individuals.

Pointers to significant others in the household are listed under the Index Terms Person Identifiers and are part of the wider group of variables listed as Key linking variables.

If you're struggling to link the files get in touch with our User Support team.

Key variables

Topic domain	Description	Variable name	In data files
Identifiers	unique cross-wave person identifier	pidp	All individual level files
	wave-specific household identifier	*w*_hidp	All wave-specific files
	wave-specific person number in the household	*w*_pno	All wave-specific individual level files
Sampling and interview variables	Individual interview outcome	*w*_ivfio	indresp, indall, indsamp
	Household interview outcome	*w*_ivfho	hhresp, hhsamp
	Individual interview mode	*w*_indmode	indresp
	Mode in which household grid was completed	*w*_modetype	indall

	Mode for the completion of the household interview	*w*_hholdmodedv	hhresp
	Primary Sampling Unit	*w*_psu	Most files (as psu in xwavedat, xwaveid)
	Strata	*w*_strata	Most files (as strata in xwavedat, xwaveid)
	Individual interview dates, adult & youth	*w*_intdatd_dv *w*_intdatm_dv *w*_intdaty_dv	indresp, indall, youth, child
	Household interview date	*w*_intdatey *w*_intdatem *w*_intdated	hhresp
Residence variables	Whether lives in England, Wales, Scotland or Northern Ireland	*w*_country	Most wave-specific files
	Which of the 12 UK regions in the UK lives in	*w*_gor_dv	Most wave-specific files
	Whether lives in urban or rural area, derived	*w*_urban_dv	Most wave-specific files
Socio-economic and demographic characteristics	Year of birth (derived from multiple sources)	doby_dv	Most individual level files
	Age at time of interview	*w*_age_dv	Most wave-specific individual level files
	Sex (derived from multiple sources)	*w*_sex_dv	Most individual level files
	whether born in England, Scotland, Wales, Northern Ireland or outside UK (self-reported)	ukborn	xwavedat
	whether born in the UK or outside UK (derived from multiple sources)	bornuk_dv	xwavedat
	ethnic group (self-reported)	racel_dv	xwavedat, indresp, indall
	ethnic group - derived from multiple sources	ethn_dv	xwavedat, indresp, indall
	De-facto marital status	w_mastat_dv	indresp, indall
	Harmonised de facto marital status	w_marstat_dv	indresp, indall

	No. of own (biological, adopted, step) children under the age of 16 in the household	*w*_nchild_dv	indresp, indall
	Highest qualification status	*w*_hiqual_dv	indresp
	Highest academic qualifications	w_qfhigh_dv	indresp
	School leaving age	scend_dv	xwavedat
	Current economic activity status	*w*_jbstat	indresp
	Socio-economic Classification of current job, SOC 2000 3-digit	*w*_jbsoc00_cc	indresp
	Socio-economic Classification of current job, NS-SEC Eight/Five/Three Class	*w*_jbnssec8_dv *w*_jbnssec5_dv *w*_jbnssec3_dv	indresp
Health & wellbeing	General health	*w*_sf1 *w*_scsf1 (response from these should be combined)	indresp
	SF-12: mental health component score	*w*_sf12mcs_dv	indresp
	SF-12: physical health component score	*w*_sf12pcs_dv	indresp
	long-standing illness or disability status	*w*_health	indresp
	subjective wellbeing (GHQ): Likert status	*w*_scghq1_dv	indresp
	subjective wellbeing (GHQ): Caseness status	*w*_scghq2_dv	indresp
	Overall life satisfaction	*w*_sclfsato	indresp
Individual and family background	Standard Socio-economic Classification (SOC 2000) of first job after leaving full-time education. Condensed three-digit version	j1soc00_cc	xwavedat
	mother's ethnic group	maid	xwavedat
	mother's country of birth	macob	xwavedat

	mother's educational qualification when respondent was aged 14	maedqf	xwavedat
	Standard Occupational Classification 1990/2000/2010 of mother's job when respondent was aged 14	masoc90_cc masoc00_cc masoc10_cc	xwavedat
	father's ethnic group	paid	xwavedat
	father's country of birth	pacob	xwavedat
	father's educational qualification when respondent was aged 14	paedqf	xwavedat
	Standard Occupational Classification 1990/2000/2010 of father's job when respondent was aged 14	pasoc90_cc pasoc00_cc pasoc10_cc	xwavedat
Household-level characteristics	Number of individuals in the household	*w*_hhsize	hhresp
	No. of children aged under 16 in the household	*w*_nkids_dv	hhresp
	Household composition	*w*_hhtype_dv	hhresp
	Housing tenure	*w*_tenure_dv	hhresp
For individual and household income variables **see the Income section**			
Variables without a wave prefix can be merged in from the **xwavedat** data file using **pidp**			
Household-level characteristics variables are from the household data file and can be merged in using **w_hidp**			

Derived variables

Derived variables are variables that are computed from one or more variables. Some are computed during the interview to control the routing within the questionnaire and appear in the context of the relevant module. Others are computed post-field for the purpose of analysis and are positioned last in the data files so they can be easily identified (derived income variables are discussed in the Derived Income variables section).

Some derived variables flag whether or not a certain characteristic is true for a study member. **w_jbft_dv** is a flag for whether or not a respondent has a full-time job and **w_nemp_dv** counts the number of employed people in the household, while **w_mnpid** points to the cross-wave person identifier of the respondent’s biological mother.

A data file may have alternative versions of a derived variable, such as those which point to others in the household, for example, **w_hgbiom** and **w_mnpno** for the person number of the respondent's biological mother in the household. While **w_hgbiom** has been computed based on information collected during the interview, **w_mpno** has been computed post-field after the information collected in the household grid has undergone extensive data cleaning.

Variables that are produced post-field, are clearly marked in the data by suffixes: UKHLS weights are shown by the suffixes "**_lw**" or "**_xw**"; most derived variables are shown by the suffix "**_dv**", and pointers to other members in the household typically end on "**pno**" or "**pid**". All variables ending on "**pid**" contain the UKHLS person identifier **pidp**, not the original BHPS person identifier.

Information collected using dependent interviewing is merged with the respective information collected using independent interviewing (e.g., when a respondent did not provide the information in the previous interview, or when they are new to the Study) and stored in the data file under the variable name used for the latter). See, for example, marital status (**w_mstatsam**).

We use lookup files between SOC 2010 and other SOC versions (i.e., 2000, 2020) to derive variables corresponding to each version. Users may apply to access the Special Licence version of Understanding Society to access non-condensed versions of these codes.

Tips for analysts:

Information about how a derived variable is produced is shown in the Derived Variable Note field of the variable. The Variable Search provides descriptive statistics for each variable and, in the Origin field, lists the variables used in the computation of the derived variable. For variables that were computed during the interview, additional information is available in the questionnaires.

Analysts can also search for the description of the derived variables under the Index Term "Derived variables" on the website.

Index of Multiple Deprivation (IMD) variables

Variables containing quintiles of the Index of Multiple Deprivation (IMD) ranking for each household's LSOA 2011 area (or equivalent for Scotland and Northern Ireland) are provided in the **indall** data file for each wave of Understanding Society, onwards Wave 12, as follows:

- England - ***w*_imd2015qe_dv** (IMD 2015 version) and ***w*_imd2019qe_dv** (IMD 2019 version)
- Northern Ireland - ***w*_imd2017qni_dv** (IMD 2017 version)
- Scotland - ***w*_imd2016qs_dv** (IMD 2016 version) and ***w*_imd2020qs_dv** (IMD 2020 version)
- Wales - ***w*_imd2014qw_dv** (IMD 2014 version) and ***w*_imd2019qw_dv** (IMD 2019 version)

These have all be derived from the latest official data released by each country's statistical body. Researchers requiring access to the full indexes or to the individual indices can do so by applying for the Special Licence dataset SN 7248 (LSOA2011) or SN 6670 (LSOA2001) as appropriate. This can then be linked to the required IMD dataset that is available as open access data published by each country.

Additional datasets

The data files released under SN6614 and listed on the Data files and their descriptions page, represent the data collected in the main survey interviews. In addition to this, data is collected from Understanding Society survey participants for specific purposes in separate surveys: Nurse health assessment and genetics data and the Covid-19 survey. These data can be linked to the main survey data using the unique cross-wave individual identifier, **pidp**.

Information about current and past histories on specific events such as employment, jobs, and (marital) partnerships is collected during the main survey interviews and the aim is to collate these and provide event history data files. Currently the partnership history file is available, the others will be made available in the future. These data can be linked to the main survey data using the unique cross-wave individual identifier, **pidp**.

Data linkage

Data linkage is the activity of bringing together separate data sources by identifying and matching the same entity in each and then bringing those different sources of information together into a single dataset. Understanding Society survey data can be linked to many different external data sources including Census data, area level deprivation, National Pupil database and so on. To learn more about these different datasets and how to access them please see the Data Linkage pages.

Interviewer characteristics

Some basic demographic information (sex, ethnicity, years of experience, age) about Understanding Society interviewers is provided by the fieldwork agency and this information can be linked to the main survey data using (scrambled) interview IDs using Study number 8579. In addition to this, Wave 1 interviewers participated in a survey which collected information on more subjective measures, attitudes and opinions. You can find this data under Study number 7615.

Comparisons of calendar year data

A wave consists of a 24-monthly samples with participants interviewed at regular 1-year intervals. As some samples are fielded in the first 12 months (BHPS and the GPS-NI, GPS2-NI samples, some in months 13-24 (IEMB sample) and some across all 24 months (GPS-GB, GPS2-GB samples and the EMB sample), just using data from the same wave to compare the two consecutive years will result in comparing different samples (see the Study design for details on each sample). Similarly, just using data from year 1 or year 2 of a wave to conduct cross-sectional analyses of that year will result in analysing samples that are not-representative. So, to correctly do these types of analyses, data from two waves need to be combined. For example, for 2019, use data from year 2 of Wave 10 and year 1 of Wave 11.

To make this process easier we have created a ready prepared calendar year dataset containing data for a whole year under a separate study number starting with the calendar year dataset for 2020 released in early 2022. A user guide accompanies the dataset which contains data from the second year of data collection for Wave 11 and data collected in the first 12 months of fieldwork for Wave 12. This dataset is not intended for longitudinal use and is for cross-sectional use only. It contains core questions but also the rotating modules (i.e., modules only asked in Wave 11 and some only asked in Wave 12). Note the 2019 calendar year data have been released with the COVID-19 Survey data and contain the second year of Wave 10 and first year of Wave 11.

The new cross-sectional calendar year datasets are planned for each subsequent year and will contain data from the second year of one Wave and early release of the first year of the next Wave. For example data from the year 2022 would contain data from the second year of data collection for Wave 13 and (at the time) early release of the data already collected in the first 12 months of fieldwork for Wave 14.

If you would like to analyse yearly data for years prior to 2019, you can produce calendar year datasets by combining data collected in a specific year from all relevant waves which were being fielded in that year. For example, for the 2012 dataset, combine data from Waves 2, 3 & 4 by selecting cases who were interviewed in 2012.

Connecting family members and households over time using the family matrix xhhrel file

Making family connections across different households over time can be difficult. To help researchers link families we have created a family matrix data file (**xhhrel)**. This data file identifies all family relationships across the Study, allowing researchers to see where people and households connect.

Each year information is collected about core sample members and who they are currently living with in the same household, every year from the time they are selected into the Study. This is provided in the file **w_egoalt** and includes the relationships of every household member i.e., how each household member pair relate to each other. However, this only records information about co-resident household members. Combining information from across the waves, we can attach information about family members to the respondent (even if they are not co-resident). This enables analysts to expand the type of research questions they are able to answer. For example, do individuals tend to partner with those with similar educational qualifications? What is the degree of social mobility in the UK? To help users link families more easily the family matrix (**xhhrel**) identifies all family relationships consistently across the study and helps facilitate research on family connections.

The **xhhrel** file creates an individual level cross-wave file of all sample members (those who were ever enumerated as part of the study) that contains familial relationship identifiers reported over the survey period for each sample member. This file also contains an origin household identifier variable (**osm_hh**), which identifies the household they come from, so that sample members who are connected with each other can be identified (either because they were co-resident at some point or were co-resident with individuals who were co-resident with each other).

Further details about why we should use the family matrix **xhhrel** file, its data structure and linking to the main survey can be found in the Family Matrix (**xhhrel)** User Guide. To help use the file, we have provided syntax for some common procedures we think users might want to do.

International Comparisons

Understanding Society is part of a world-wide family of household panel studies the Household Income and Labour Dynamics in Australia (HILDA), the Korea Labor and Income Panel Study (KLIPS), the Panel Study of Income Dynamics (PSID), the Russia Longitudinal Monitoring Survey (RLMS-HSE), the Swiss Household Panel (SHP), the Canadian Survey of Labour and Income Dynamics (SLID), and the German Socio-Economic Panel (SOEP). Information about these can be found on our website. The Comparative Panel File is a user-generated syntax is available that allows harmonisation of data from seven household panel surveys.

Using the British Household Panel Survey (BHPS)

The British Household Panel Survey began in 1991 and is the base on which Understanding Society is built.

The Wave 1 panel consisted of around 5,500 households and 10,300 individuals drawn from 250 areas of Great Britain. Additional samples of 1,500 households in each of Scotland and Wales were added to the main sample in 1999, and in 2001 a sample of 2,000 households was added in Northern Ireland, making the panel suitable for UK-wide research. As this was a longitudinal household study, households were sampled and then individuals living in these households were followed wherever they moved to within the UK. All 16+ year old household members were eligible for adult interviews, while from Wave 4 onwards, all 11-15 year olds were eligible for youth self-completion questionnaires. Interviews were conducted face-to-face with a few by proxy interviews and a few by telephone.

As part of wave 18, BHPS participants were asked if they would consider joining the new, larger and more wide-ranging survey Understanding Society. Almost 6,700 of just over 8,000 BHPS participants invited to join did so. The continuing sample from the British Household Panel Survey (BHPS) joined the Understanding Society sample in Wave 2. The cases in the two samples can be distinguished using the variable **w_hhorig**. The variable also allows the identification of different components of the BHPS sample (see below). This will allow continued research on changes in people's lives from 1991.

The sample design, data structure, interview pattern, following rules are very similar across the two surveys. The core questions included in BHPS were included in the new Study to allow researchers to study trends and changes in people's lives starting from 1991. There are, however, a few differences in the data collected. New topic areas were included such as more questions related to health and wellbeing, ethnicity and migration, social networks, and topical issues such as the Olympics and EU referendum.

Harmonised or stand-alone BHPS

To use the long run of data collected from BHPS sample members along with the data collected as part of Understanding Society, users are advised to use the harmonised BHPS, which is included in the Understanding Society data release. If using the stand-alone BHPS (SN5151) then they will have to undertake the harmonisation exercise themselves.

Both options facilitate linking cases across studies using the unique UKHLS person identifier **pidp** which has been added for all BHPS sample members in the BHPS data files (i.e., irrespective of whether they have ever participated in the UKHLS).

There are advantages in using the harmonised BHPS files: all variables that are equivalent in both studies have been renamed so they have the UKHLS name and efforts have been made to assure that the information content is identical. If the name of a variable was the same but the content was not identical then the BHPS variable received the suffix **_bh**. All BHPS variables with names that do not correspond to any variable in UKHLS are left as is.

The same convention was followed for data files. The **xwavedat** files from both studies have been merged and all information that is available for cases from both samples has been harmonised. In this harmonised version, all BHPS Wave 1-18 file names and variable names have a prefix "**bw_**".

Harmonised BHPS data are documented in the Understanding Society online data documentation, including the variable occurrence stretching across both studies removing the absolute requirement to jump across study documentations.

Users should note that the harmonisation project is ongoing and a number of data aspects that could be harmonised in principle have not yet been harmonised due to the complexity of the task and time constraints. More detail about the harmonised BHPS is provided in the designated Understanding Society harmonised BHPS User Guide.

If using the stand-alone, non-harmonised, BHPS data (SN5151) it is important to be aware that variable names in the BHPS dataset have slightly different formats:

- They are limited to eight characters
- There is no underscore separating the wave prefix from the main part of the name, so the **indresp** file for Wave 1 is named **aindrep** and sex variable for Wave 1 is named **asex**.
- Derived variables, imputation flags, weights and other special variables are not distinguished by "**_dv**" or "**_if**" suffixes
- As this is the stand-alone non-harmonised version variables with same names as in the UKHLS files may not have the same meaning

Tips for analysts:

Both samples can be used for cross-sectional and longitudinal analyses. For the appropriate weights in the harmonised version refer to the Weighting Section. For the stand-alone version (SN 5151) refer to the Weighting Section V. A5-1 in the BHPS user guide.

If you want to use the harmonised BHPS across the 30+ years of data available in the BHPS sample, you need to consider how to treat Wave 2 of Understanding Society. As the BHPS sample was included in Wave 2 of Understanding Society, you could treat that as Wave 19. But the interviews in this wave were conducted from January to December 2010 while interviews for BHPS Wave 18 were conducted from September 2008 to December 2009 the interval between these two interviews may not be around 12 months for everyone.

Questionnaire Content

Understanding Society covers a wide range of topics: education, employment, family and households, income, Health and wellbeing, finance, housing, expenditure and deprivation, politics and social attitudes.

The survey also collects demographic information (age, sex, ethnic group, country of birth), family background information (e.g., parents' occupations, parents' grand parents' countries of birth), attitudes and values (ethnic identity, political beliefs and support, gender attitudes etc).

A core set of questions are asked every year, while some are asked every few years as these are not expected to change every year. In addition to these there are questions that are asked only if the respondent has experienced a particular event such as childbirth, or reached the age of 45, and so on. The long-term content plan summarises the pattern that has been collected.

Health and Biomarkers

Understanding Society asks a wide range of questions on health, wellbeing and biomarkers. As part of the main study, from Wave 1 onwards participants are asked a number of questions about their general health (physical and mental), healthy & unhealthy behaviours, hospitalisations and so on. In Wave 2 and Wave 3 adult participants received a follow-up health assessment visit from a registered nurse. A range of bio-medical measures were collected from around 20,000 adults, which included blood pressure, weight, height, waist measurement, body fat, grip strength and lung function. Blood samples were also taken at these visits and biomarker data is available including cholesterol and triglycerides, Glycated haemoglobin HbA1c, and so on. Further details on using this health data and genetics and epigenetics data are on the Understanding Society website.

Ethnicity and immigration

The Ethnic Minority Boost Sample (EMBS) and Immigrant and Ethnic Minority Boost Sample (IEMBS) were introduced at Waves 1 and 6 to allow research on different ethnic groups and immigrants. While all participants answer the same basic questionnaire, a subsample comprising mainly ethnic minorities and immigrants living in Great Britain are asked an additional five minutes of questions along with a comparison sample of 500 households from the general population. These are questions of particular relevance to research on ethnicity and migration. Note, the basic questionnaire also includes ethnicity-related questions about issues such as ethnic group, national identity, own, parents' and grandparents' countries of birth and childhood language. Visit the long term content plan for details and explore the Ethnicity and immigration research User Guide.

Questionnaires, questionnaire modules and Index Terms

Questionnaires

The PDF versions of the questionnaires can be found on the documentation pages of the Understanding Society website. These include the household grid, household questionnaire, adult individual, self-completion and proxy questionnaires, and are an important source of information about the wording of individual questions, who was asked, and what questions precede and follow.

Most of the interview is conducted with a computer-assisted personal interview (CAPI) with a few conducted over the telephone, Computer-assisted telephone interview (CATI), and from Wave 8 onwards also via computer-assisted web interview (CAWI). These instruments govern the flow of questions and recording of answers. The text within the PDF questionnaires can be searched for specific words, such as variable names or words in questions. These pdf files include all the different questionnaires except for the youth questionnaires (all waves) and adult self-completion questionnaires for Waves 1 and 2. The PDF files for the adult self-completion questionnaires for Waves 1 and 2 as well as the youth questionnaires (all waves) correspond to the way they appeared to participants (with the addition of annotated variable names which were added later).

These are also included in the zipped folder when you download the data from the UK Data Service.

Questionnaire modules

The principal adult questionnaires are organised into modules. Modules can be searched for in the online documentation system. About half of the questionnaire content is collected annually, with additional modules collected at different intervals, often every two, three or four years. The paper self-completion questionnaires carried at Waves 1 and 2 were not divided into modules as these were separate standalone documents. From Wave 3 onward, the self-completion content was carried as CASI modules, where the interviewer would turn the laptop towards the participant who would answer the questions using the laptop by themselves.

Translated questionnaires

Instruments and survey materials for Waves 1 -7 were translated into multiple languages: Welsh, Arabic, Bengali, Cantonese, Gujarati, Punjabi (in Urdu and Gurmukhi scripts), Somali and Urdu. Translated documents can be requested by email from info@understandingsociety.ac.uk. For the new Immigrant and Ethnic Minority Boost, introduced in Wave 6, instruments and survey materials were also translated into Polish, Portuguese and Turkish. From wave 8 onwards the survey instruments and materials were translated into Welsh, Bengali, Gujarati, Punjabi (in Urdu and Gurmukhi scripts), Urdu, Polish, Portuguese and Turkish for everyone. For the first time at Wave 13, the web version of the questionnaire was also translated, so translated interviews could be conducted in any mode.

Index terms

All variables have been linked to one or more Index Terms. This makes it easier to find variables. For example, if you are interested in researching childcare you can click on the Index Term “childcare” and see all the variables associated with it.

Finding variables

Understanding society is a vast survey including thousands of questions and variables associated with them. Find the variables you need for your research by using the Variable search facility. Search by variable name, by data file or by index term or question module. This facility provides links between questions, variables and data files. Explore the video How to use the Variable search. Our list of key variables for the analysis of individual response data may also help you.

Creating syntax with the Code creator

To help researchers get started with their research the Code creator extracts data from different data files in the Main Study and produces a simple flat data file in either wide of long format. Researchers can select the variables needed from the Variable Search, ‘save’ the variables and ‘build’ the code. It provides researchers with ready-to-use Stata syntax to run on the downloaded data.

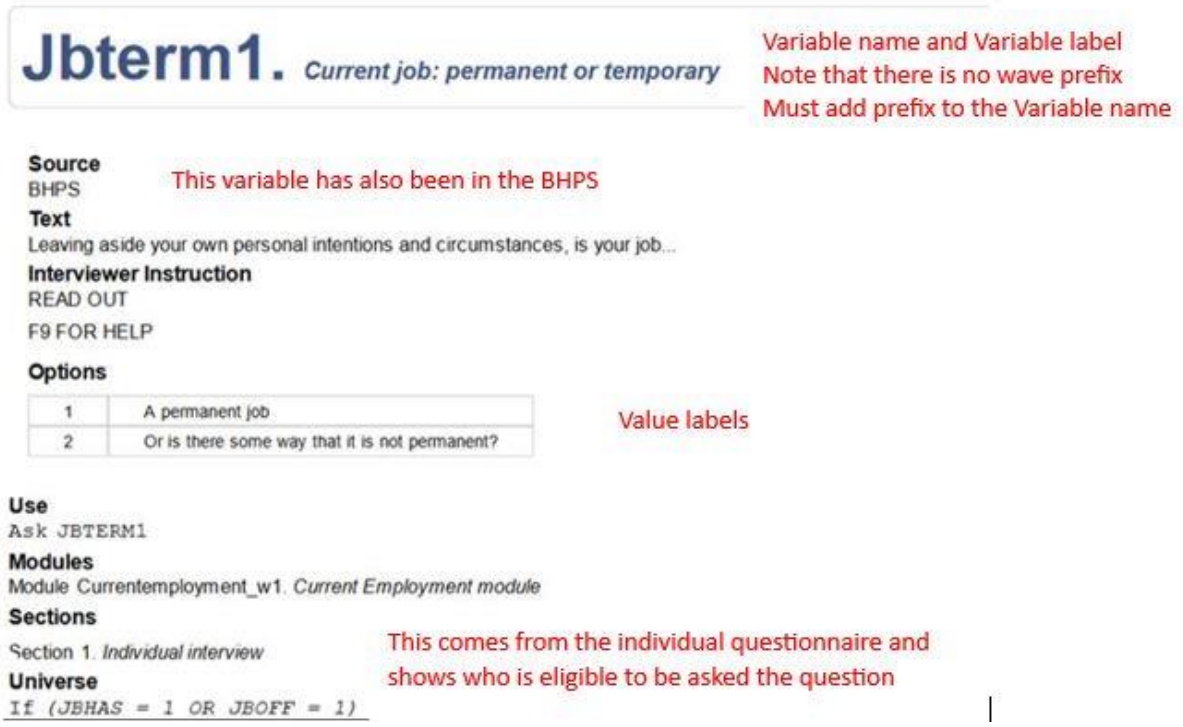
Reading the Questionnaire

Find the variables you need for your research by using the variable search facility. Search by variable name, by data file or by index term. This facility provides links between questions, variables and data files. Our list of key variables for the analysis of individual response data may also help you.

The key to understanding the Study is to read the questionnaires. It is complex but the benefit of reading the questionnaire is that it will help you understand who is eligible for a question, which questions follow and precede that question, the structure of the data and its use.

The questionnaire tells you what to expect, for example, it will tell you why there are missing cases and why some questions return multiple variables. These are explained with some examples below. Note the variable names in the questionnaire do not contain the wave prefix that is applied in the data files.

Example of a question that is not asked of everyone



Jbterm1. *Current job: permanent or temporary*

Variable name and Variable label
Note that there is no wave prefix
Must add prefix to the Variable name

Source
BHPS

This variable has also been in the BHPS

Text
Leaving aside your own personal intentions and circumstances, is your job...

Interviewer Instruction
READ OUT

F9 FOR HELP

Options

1	A permanent job
2	Or is there some way that it is not permanent?

Value labels

Use
Ask JBTERM1

Modules
Module Currentemployment_w1. *Current Employment module*

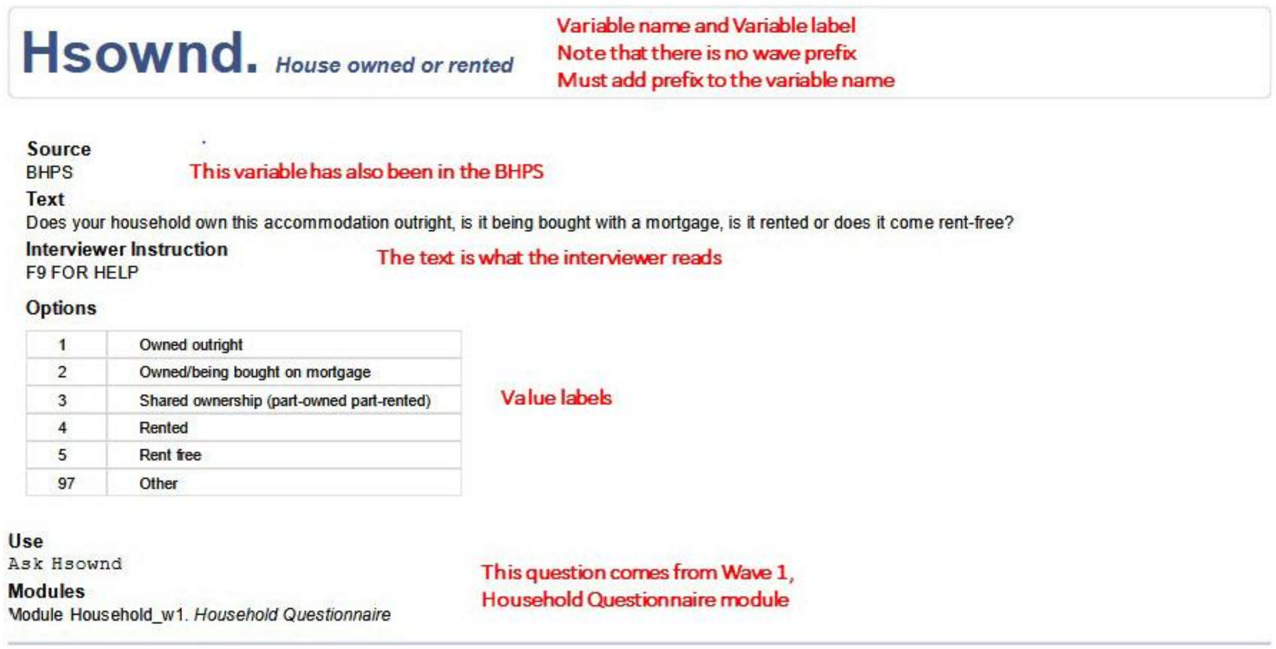
Sections
Section 1. *Individual interview*

Universe
If (JBHAS = 1 OR JBOFF = 1)

This comes from the individual questionnaire and shows who is eligible to be asked the question

All questions are not asked of everyone if they are not relevant. In this example, the question about whether the current job is permanent or temporary is only asked of those who say they have a job. These eligibility rules are detailed in the Universe field below the question. This video gives an overview of questionnaire routing and missing values in Understanding Society.

Example of a question in the household questionnaire (this variable is also in the BHPS)



Hsownd. *House owned or rented*

Variable name and Variable label
Note that there is no wave prefix
Must add prefix to the variable name

Source
BHPS

This variable has also been in the BHPS

Text
Does your household own this accommodation outright, is it being bought with a mortgage, is it rented or does it come rent-free?

Interviewer Instruction
F9 FOR HELP

The text is what the interviewer reads

Options

1	Owned outright
2	Owned/being bought on mortgage
3	Shared ownership (part-owned part-rented)
4	Rented
5	Rent free
97	Other

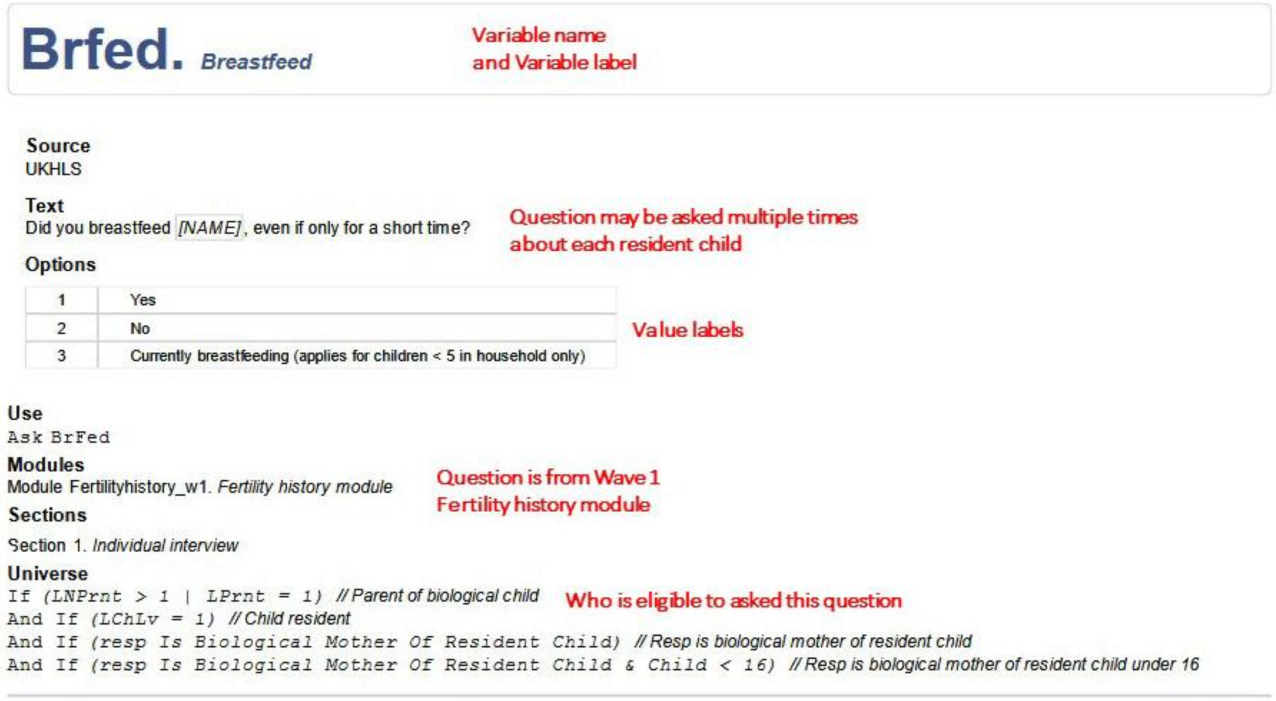
Value labels

Use
```
Ask Hsownd
```

Modules
Module Household_w1. *Household Questionnaire*

This question comes from Wave 1,
Household Questionnaire module

Example of question with looping from individual questionnaire



Brfed. *Breastfeed*

Variable name
and Variable label

Source
UKHLS

Text
Did you breastfeed *[NAME]*, even if only for a short time?

Question may be asked multiple times
about each resident child

Options

1	Yes
2	No
3	Currently breastfeeding (applies for children < 5 in household only)

Value labels

Use
```
Ask BrFed
```

Modules
Module Fertilityhistory_w1. *Fertility history module*

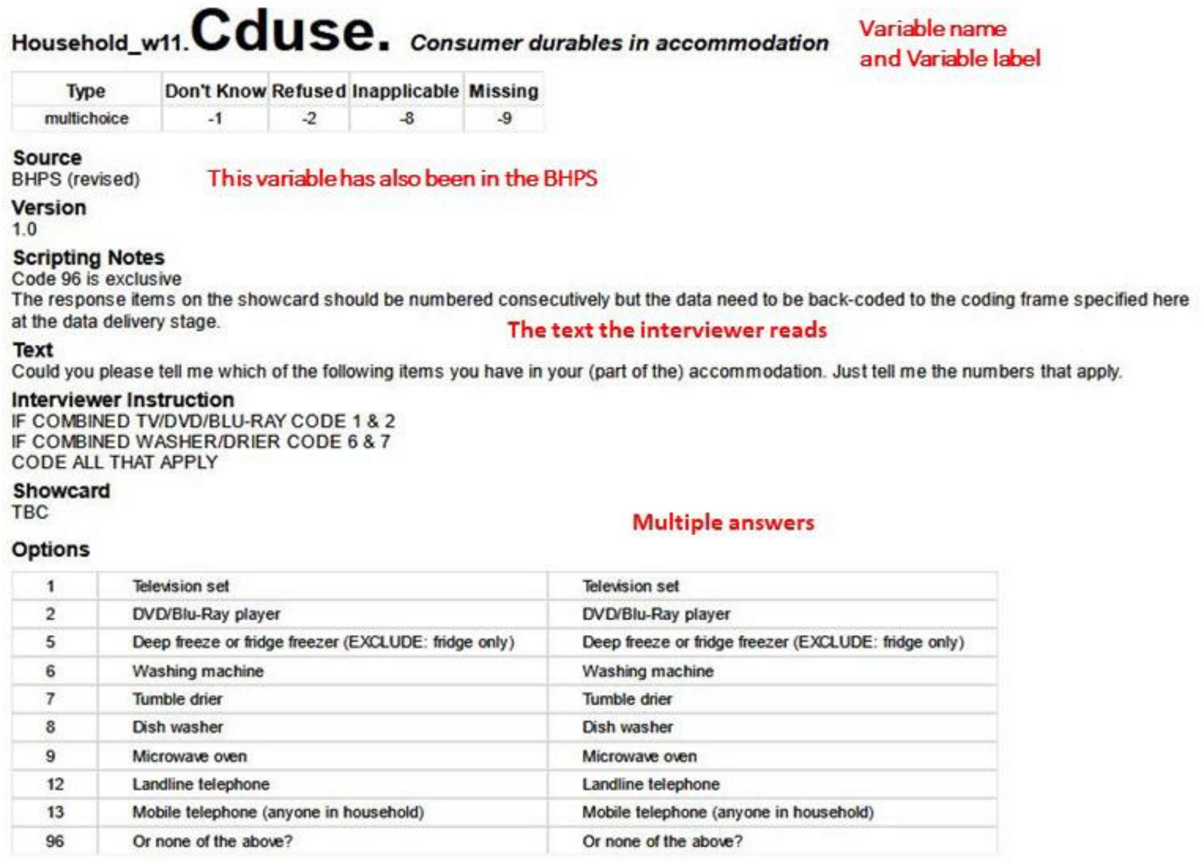
Question is from Wave 1
Fertility history module

Sections
Section 1. *Individual interview*

Universe
```
If (LNPrnt > 1 | LPrnt = 1) // Parent of biological child
And If (LChLv = 1) // Child resident
And If (resp Is Biological Mother Of Resident Child) // Resp is biological mother of resident child
And If (resp Is Biological Mother Of Resident Child & Child < 16) // Resp is biological mother of resident child under 16
```

Who is eligible to asked this question

Example of question with multiple choices resulting in multiple variables



Household_w11.**Cduse.** *Consumer durables in accommodation*

Variable name and Variable label

Type	Don't Know	Refused	Inapplicable	Missing
multichoice	-1	-2	-8	-9

Source
BHPS (revised)

This variable has also been in the BHPS

Version
1.0

Scripting Notes
Code 96 is exclusive
The response items on the showcard should be numbered consecutively but the data need to be back-coded to the coding frame specified here at the data delivery stage.

The text the interviewer reads

Text
Could you please tell me which of the following items you have in your (part of the) accommodation. Just tell me the numbers that apply.

Interviewer Instruction
IF COMBINED TV/DVD/BLU-RAY CODE 1 & 2
IF COMBINED WASHER/DRIER CODE 6 & 7
CODE ALL THAT APPLY

Showcard
TBC

Multiple answers

Options

1	Television set	Television set
2	DVD/Blu-Ray player	DVD/Blu-Ray player
5	Deep freeze or fridge freezer (EXCLUDE: fridge only)	Deep freeze or fridge freezer (EXCLUDE: fridge only)
6	Washing machine	Washing machine
7	Tumble drier	Tumble drier
8	Dish washer	Dish washer
9	Microwave oven	Microwave oven
12	Landline telephone	Landline telephone
13	Mobile telephone (anyone in household)	Mobile telephone (anyone in household)
96	Or none of the above?	Or none of the above?

Questions that are not asked every year

Each year we ask respondents a set of core questions, but questions which ask about aspects of people's lives that are not likely to change frequently, are asked less often. This is done to make sure the interviews are not very long (so reduce respondent burden) and that the questions are meaningful and relevant. For example, it is not sensible to ask someone's country of birth every year as that will not change over time.

Initial conditions questions

Initial condition questions are asked only once, generally the first time someone is interviewed as these are not expected to change, e.g., country of birth, date of birth. These questions may not be asked in Wave 1 for everyone in the sample, as not everyone joins the survey in its first wave. Those who join the household after the first wave will be asked in the wave they join. Similarly, when new samples are added these questions are first asked in that wave, e.g., these initial conditions questions were asked of the IEMB sample in Wave 6 (2014-2016) when the sample was first introduced. To make it easier to use, we have gathered the answers to these initial conditions questions, in whichever wave they were answered, and put them together in one individual level file **xwavedat**.

Rotating modules

Rotating modules refer to questions that are asked every few years as they are expected to change less frequently, such as a respondent's wealth, friendship network, personality traits and so on. This is done to reduce respondent burden. To fill data gaps between the rotations, some researchers and analysts use a respondent's data from the nearest rotation. However, this approach may not be appropriate in all research applications and is left for the researcher to determine.

Event triggered questions

Some questions are triggered by changes in individual circumstances. For example, questions about retirement planning such as expected age of retirement (**ageret**) are asked when the respondent is aged 40, 45, 50, 55, 60 or 65 and still not retired.

Specific questions about one-off global events are added to the questionnaire in the relevant waves to expand the study's research possibilities. For example, questions about the 2012 London Olympics (in Wave 4 (2012-2014) and Wave 5 (2013-2015)), or experiences during the Covid19 pandemic (such as Covid19 symptoms **Pcovidsymp** or the amount of furlough payment received **Jsseissam** in Wave 11 (2019-20) and Wave 12 (2020-21)).

Other questions are triggered by specific global events, such as questions about the voting behaviour during the last General Elections (**Vote8**) are asked of respondents interviewed in the months after the General Elections (e.g., Waves 11 and 12) following the general election in December 2019, Waves 9 and 10 during June 2017 - May 2018 following the general election in June 2017).

Changes to the Questionnaire and questions

We collect information from our sample members each year to build a picture of how society is changing over time or remaining the same. By asking the same questions of the same set of people each year it allows researchers to measure change over time. However, there are several reasons why a question or its response options may change over time or why a question may not be asked every year.

Changes to questionnaire content are never made lightly and content is only updated if it is of clear benefit to the Study and enriches the data for research. When questions are reviewed, relevant topic experts and data users are consulted on proposed changes and ensure longitudinal compatibility for researchers.

Questions included in the questionnaires/survey instruments undergo thorough testing and validation and are based on robust methodological research.

Questions from other surveys are used where appropriate. All survey instruments/questionnaires are tested so that any issues with question wording and routing, interview flow and timings can be identified, and appropriate changes made before the survey is implemented or fielded.

Question changes to reflect societal change

As society changes, some questions need reviewing to align with current practices or changes in policy. For example since the Study began the way we work has changed and reviewing these questions means we can incorporate those changes. To capture information on respondents current employment and those with multiple jobs, changes to the questionnaire were first made in Wave 13 to the **Currentemployment** module. Previously questions had focussed on the assumption that people have one paid job with a ‘standard’ working week. These questions were redesigned to cater for respondents with multiple jobs taking into account non-standard working, including portfolio working, where people have lots of smaller jobs rather than one main job.

Some questions become unnecessary or obsolete such as whether you have a ‘colour television’ or ‘video recorder/DVD player’ **Cduse**. These were replaced with ‘television set’ and ‘DVD/Blu-Ray player’ respectively onwards from Wave 8 (2016-2018).

This can also work in reverse with the introduction of previously non-existent items such as iPads, electric cars and changes in behaviour such as meeting friends online and owning smartphones **smartmob**. Variable **Mobcomp** introduced in Wave 5 is now a core question asked every year asking if respondents have a tablet or iPad. To reflect changes in how young people feel and interact socially, changes in questions asked to 10-15 year olds were introduced in Wave 12 asking how often they get together with friends in person or online **Ypfndonl,** as well as a question about how often they feel lonely **Yplonely**. New questions can respond to an increase in awareness and inclusivity within society, such as how participants self-identify their sex and gender and how this might have changed from that described at birth. Wave 12 variables **Birthsex** and **Genderself** were asked of young adults aged 16-21 in the self-completion module **Scasexandgender**. Questions can also respond to changes in the economy such as the questions added in Wave 11 to measure the experience of gig economy type of jobs (see module **Gig economy**).

Changes in policy

Changes in policy can result in new response options. For example, within **Jbstat**, following the introduction of shared parental and adoption leave. In Wave 13 the Study introduces more comprehensive questions about parental leave. Question and response wording also change with the introduction of new qualifications or state benefits. For example, GCSE grades changing from letters to numbers in England and Wales in 2017 (**GCSE5**). Introduced in Wave 11 **GCSEatoceq** asked new entrants to the Study the number of GCSEs or equivalent qualification names and/or grades to accommodate those devolved nations who did not change qualification name and/or grades. Wave 14 (2022-2024) sees the introduction of new options for income variable **Ficode**, asking about income from new Government payment schemes for Child, Adult and Pension age disability as well as the Scottish child payment scheme.

Changes in the political landscape can also prompt additional modules such as voting behaviour in the devolved elections in Scotland, Wales and Northern Ireland asked in Wave 12 in modules **scadevolvedScotland, Scadevolvedwales** and **Scadevolvedni**.

Enriching the data collected

Response options can respond to changes in education pathways e.g., with the addition of two new degree apprenticeship options in variable **Apprent** in Wave 12**.** Similarly, responses can be aligned with source data to enable comparisons with other studies, such as the extra questions aligned with the Gender and Generation survey for child responsibilities in the **Domesticlabour** module in Wave 12.

Additional questions can also enrich the data in understanding people’s income. Wave 12 captured payments for student loans within the household income for the first time in the **Studentloans** module. Altering the universe can enrich the data too: those without a job are asked if they were looking for a job in the last 4 weeks (**julk4**). From onwards, Wave 14 (2022-2024) this question also asks those in employment as well. To define missing or inapplicable values, questions previously asked only to the extra 5 minute sample in the **Harassment** module **(attacked)** were asked to the entire universe from Wave 11 onwards.

Improving the questionnaire

Sometimes changes are made to the questionnaire across waves when routing errors only become known after data collection has been completed. (e.g., in Wave 1 only the proxy interview included the question for whether or not a respondent had access to a car (***w*_drive**) and from Wave 2 (2010-2012) onward this information is available for both adult and proxy respondents (those with missing information from Wave 1 were asked again).

Introduction of new modes of data collection

The switch in mode from paper self-completion to the CASI on the laptop in Wave 4 meant that for some questions the response options were presented differently between waves. For example, response options were arrayed horizontally in the paper self-completion (e.g., satisfaction questions), and vertically in CASI. There is some evidence that the change in the way the response options were presented may affect how some people respond to the question (Budd, Gilbert et al. 2012).

From Wave 7 (2015-2017) onwards an increasing proportion of the sample were offered the opportunity to complete the questionnaire via the web. To adjust for the change in mode and reduce mode effects, some question wording and structure changed for the web mode with interviewer instructions and help tips, usually read by the interviewer, included in the question text. For example, in web mode, the question ‘In whose name is this accommodation rented?’ (**Rentp**) includes the extra wording ‘This is the person responsible to the landlord for the rent. If there is a joint tenancy, record all those responsible.’ Similarly questions with showcards give further clarification. For example in the face-to-face mode, after asking the question about consumer durables in the accommodation (**Cduse**), interviewers present the showcard to respondents and ask the numbers that apply. In the web mode, the question text includes this additional clarification “If you have a combined TV/Video select ‘Colour television and 'Video recorder/DVD player'. If you have a combined washer drier select 'Washing machine' and 'Tumble Drier'.”

Within wave changes

On rare occasions question text changes within a wave, or questions are dropped within a wave. For example, at the end of the first six months of data collection in Wave 1, multiple variables were dropped because of the length of the interview to reduce the burden on respondents, e.g., employment history module, parents’ educational qualification questions. The rest of the sample were asked these questions in subsequent waves (parents’ educational qualification questions were asked in Wave 2 and employment history in Wave 5). Similarly, with the start of the Covid19 pandemic in March 2020, questions about furlough and the coronavirus symptoms were added to Wave 11. These were asked of respondents who were interviewed from onwards April 2020.

Variable notes

To help researchers identify variables where the question text may have changed over time, variable notes are being created and added to the Variable search on the Understanding Society website.

Other fieldwork materials such as showcards, advance letters and interviewer instructions are also on the website.

Tips for analysts: COVID-19

In response to the pandemic the questionnaire was adapted to capture changes during this time. Updates went into the field on 28 July 2020. The updates are described in the document Understanding Society main study changes due to the COVID-19 pandemic.

Note a separate short monthly (later bi-monthly) online Covid-19 survey was conducted from April 2020 to September 2021 to quickly collect and evaluate the effect of the pandemic on people's lives. This data can be linked to the main survey data using the unique cross-wave identifier, **pidp**.

Changes to employment questions from Wave 13

Since the BHPS and Understanding Society first began, the way we work in the UK has changed. Understanding Society's questions on employment were inherited from the BHPS and designed to measure employment at the beginning of the 1990's. To reflect societal changes in jobs and how people work, Understanding Society undertook the redevelopment of the employment content of the survey. Some of the changes made also incorporated changes in policy, such as regulations around zero-hour contracts, and the questions have been updated to reflect this. We also sought to enrich the range of employment-related data we collect to increase the range of research possible. The questions were redesigned to better cater for those respondents with multiple jobs taking into account non-standard working, including portfolio working, where people have multiple smaller jobs rather than one main job. Wave 13 of the Main Study is the first instance of changes to job related variables.

To improve the employment-related content, we held a workshop in October 2019 bringing together experts in the field, including substantive and methodological researchers. Presentations were made on the gig economy, precarious labour conditions, and the transformation of the Labour Force Survey. The discussions held throughout the day formed the basis for the redevelopment plan, and were supplemented by communications, reports, and publications relevant to the proposed updates.

Prior to Wave 13, questions had focused on the assumption that people have one paid job with a 'standard' working week. For Waves 1 to 12, data were collected on current main job and second job. From Wave 13, the questionnaire design allows up to 16 jobs to be reported (please note that the maximum number of jobs reported in Wave 13 was 10) and the "Second jobs" module is no longer asked. Please refer to the Wave 13 questionnaire for these questions.

As always, the variables on current employment are in the **indresp** file. For each job the respondent reports, they are asked a series of questions about the characteristics of that job which are also available in **indresp.** So, there is a separate set of variables for each job. The variables have the suffix "**_X**", referencing the job to which the variable belongs. For example, **jbterm_1** refers to the first job reported, **jbterm_2** to the second job reported, etc. The job-specific variables are documented in Table 6.1 (the prefix "**w_**" is a placeholder that identifies the wave of the survey). Table 6.1 also has new variables containing information that we did not previously collect (e.g., about zero hours contracts and the gig economy), referenced as 'New', and variables that we previously only collected about the 'main' job and now ask of all jobs (e.g., occupation or industry), referenced as 'UKHLS'.

The variable, **multijobs**, records the number of jobs reported, and **jbmain** records which of the jobs they reported is their main job. To facilitate longitudinal analyses with variables collected in prior waves, we have created derived variables that contain the characteristics of the current 'main' job and have the same names as in previous waves. For example, if the respondent reported the third job as their main job, so **jbmain** = 3, then the variable **jbsoc10_cc** contains the value of **jbsoc10_3_cc**. These variables are documented in Table 6.2. Note at Wave 14 the main current job occupation variables (jbsoc00, m_jbsoc10) were revised to achieve consistent measurement across waves (see Revisions to the main current job occupation in Wave 13).

Please note, that there are some questions that are still asked about the main job (which is either the only job reported or if more than one job then the job chosen as the main job in **jbmain**). These questions are in the module "Employees_w13". The "Universe" for these questions confirms this routing rule:

> If ((CURRENTEMPLOYMENT.MULTIJOBS = 1 & CURRENTEMPLOYMENT.JBSEMP) = 1 | JOBCODE Selected At CURRENTEMPLOYMENT.JBMAIN Is CURRENTEMPLOYMENT.JBSEMP = 1) // Has one job and is an employee OR Has more than 1 jobs and is an employee in the main job

This module includes variables such as **jbmngr, jbsize, jbsect, jbsectpub, jbhrs, jbot, paygl** and so on.

The "second job" was dropped, as information about all jobs is now collected as part of the "current employment" module (for the variables affected see Table 6.3). Instead, a new module about the other jobs, otherjobs_w13, was included. This included the questions listed in Table 6.4, and the suffix **_X** showed which job number the variable was referring to, where **X** ranged from 1 to 10 in Wave 13. Respondents who reported more than one job (**multijobs**>1), were asked these questions only for the jobs that were NOT their main job. For example, if someone reported two jobs and their second job as the main job, then the variables in this module that referred to the first job will have valid values, and the variables referring to all other jobs will be -8.

Additionally, new questions were asked for those seeking jobs in the "job search" module. There are four sets of variables named **julksoc90_Y julksoc00_Y julksoc10_Y** and **julksoc20_Y**. (These questions were adapted from the COVID-19 study) Respondents were able to input up to three jobs they were seeking. Therefore, there were 12 new variables for each SOC year classification. (See Table 6.5)

Finally, the variables in the "non-employment" module, that is asked of respondents who are not currently in work, were not changed (see Table 6.6 for a list of these variables).

Table 6.1. Variables collected in current employment module for each reported job (job number indexed by suffix "_X")

Variable	Variable Label	Origin
w_zerohour_X	Hours set: Job X	New
w_gigempjob_X*	Gig economy job: Job X	New
w_gigempjob2_X*	Gig economy job: Job X	New
w_gigemptyp1_X	Providing a driving or taxi service, for a fee: Job X	New
w_gigemptyp2_X	Providing delivery or courier services: Job X	New
w_gigemptyp3_X	Providing professional work, such as consultancy, legal advice, accounting: Job X	New
w_gigemptyp4_X	Providing creative or IT work, such as writing, graphic design, or web dev: Job X	New
w_gigemptyp5_X	Providing administrative work, such as data entry or click work: Job X	New
w_gigemptyp6_X	Providing skilled manual work, such as plumbing, building, electrical maint: Job X	New

Variable	**Variable Label**	**Origin**
w_gigemptyp7_X	Providing personal services, such as cleaning, moving, or DIY tasks: Job X	New
w_gigemptyp8_X	Selling good or crafts that I have made (e.g. via Etsy, etc.): Job X	New
w_gigemptyp9_X	Selling goods that I have bought to resell: Job X	New
w_gigemptyp10_X	Renting out a place (my home or another property I own) for a short-term: Job X	New
w_gigemptyp96_X	Other type of job: Job X	New
w_gigemptyp97_X	None of these types: Job X	New
w_jbsoc90_X	Current job: SOC 1990: Job X	UKHLS
w_jbsoc90_X_cc	Current job: SOC 1990 (condensed 3 digits version): Job X	UKHLS
w_jbsoc00_X	Current job: SOC 2000: Job X	UKHLS
w_jbsoc00_X_cc	Current job: SOC 2000 (condensed 3 digits version): Job X	UKHLS
w_jbsoc10_X	Current job: SOC 2010: Job X	UKHLS
w_jbsoc10_X_cc	Current job: SOC 2010 (condensed 3 digits version): Job X	UKHLS
w_jbsoc20_X	Current job: SOC 2020: Job X	UKHLS
w_jbsoc20_X_cc	Current job: SOC 2020 (condensed 3 digits version): Job X	UKHLS
w_jbsic07_X	Current job: SIC 2007: Job X	UKHLS
w_jbsic07_X_cc	Current job: SIC 2007 (condensed 2-digit version): Job X	UKHLS
w_jbseg_X_dv	Current job: Socio-economic Group: Job X	UKHLS
w_jbrgsc_X_dv	Current job: Registrar General's Social Class: Job X	UKHLS
w_jbnssec_X_dv	Current job: NS-SEC: Job X	UKHLS
w_jbnssec8_X_dv	Current job: Eight Class NS-SEC: Job X	UKHLS
w_jbnssec5_X_dv	Current job: Five Class NS-SEC: Job X	UKHLS
w_jbnssec3_X_dv	Current job: Three Class NS-SEC: Job X	UKHLS
w_jbisco88_X	Current job: International Classification of Occupations 1988: Job X	UKHLS
w_jbisco88_X_cc	Current job: ISCO88 (condensed 3 digits version): Job X	UKHLS
w_jbterm1_X	Current job: permanent or temporary: Job X	UKHLS
w_jbterm2_X	Type of non-permanent job: Job X	UKHLS
w_jbsemp_X	Employed or self-employed for current job(s): Job X	UKHLS
w_jbbgd_X	Day started current job: Job X	UKHLS
w_jbbgm_X	Month started current job: Job X	UKHLS
w_jbbgy_X	Year started current job: Job X	UKHLS

Note: X can go from 1 to 16. In Wave 13, X is from 1 to 10. *__gigempjob__ was changed mid-field to **gigempjob2**.

Table 6.2: Characteristics of the 'main' job derived from the job-specific variables listed in Table 6.1 and the variable jbmain

Variable	Variable label	Origin
w_zerohour	Hours set	New
w_gigempjob	Gig economy job	New
w_gigemptyp1	Providing a driving or taxi service, for a fee	New
w_gigemptyp2	Providing delivery or courier services	New
w_gigemptyp3	Providing professional work, such as consultancy, legal advice, accounting	New
w_gigemptyp4	Providing creative or IT work, such as writing, graphic design, or web dev	New
w_gigemptyp5	Providing administrative work, such as data entry or click work	New
w_gigemptyp6	Providing skilled manual work, such as plumbing, building, electrical maint	New
w_gigemptyp7	Providing personal services, such as cleaning, moving, or DIY tasks'	New
w_gigemptyp8	Selling good or crafts that I have made (e.g. via Etsy, etc.)	New
w_gigemptyp9	Selling goods that I have bought to resell	New
w_gigemptyp10	Renting out a place (my home or another property I own) for a short-term	New
w_gigemptyp97	Other type of job	New
w_gigemptyp96	None of these types	New
w_jbsoc90	Current job: SOC 1990	UKHLS
w_jbsoc90_cc	Current job: SOC 1900 (condensed 3 digits version)	UKHLS
w_jbsoc00	Current job: SOC 2000	UKHLS
w_jbsoc00_cc	Current job: SOC 2000 (condensed 3 digits version)	UKHLS
w_jbsoc10	Current job: SOC 2010	UKHLS
w_jbsoc10_cc	Current job: SOC 2010 (condensed 3 digits version)	UKHLS
w_jbsoc20	Current job: SOC 2020	UKHLS
w_jbsoc20_cc	Current job: SOC 2020 (condensed 3 digits version)	UKHLS
w_jbsic07	Current job: SIC 2007	UKHLS
w_jbsic07_cc	Current job: SIC 2007 (condensed 2-digit version)	UKHLS
w_jbisco88	Current job: International Classification of Occupations 1988	UKHLS
w_jbisco88_cc	Current job: ISCO88 (condensed 3 digits version)	UKHLS
w_jbseg_dv	Current job: Socio-economic Group	UKHLS
w_jbrgsc_dv	Current job: Registrar General's Social Class	UKHLS
w_jbnssec_dv	Current job: NS-SEC	UKHLS
w_jbnssec8_dv	Current job: Eight Class NS-SEC	UKHLS
w_jbnssec5_dv	Current job: Five Class NS-SEC	UKHLS
w_jbnssec3_dv	Current job: Three Class NS-SEC	UKHLS
w_jbterm1	Current job: permanent or temporary	UKHLS
w_jbterm2	Type of non-permanent job	UKHLS
w_jbsemp	Employed or self-employed for current job(s)	UKHLS

w_jbbgd	Day started current job	UKHLS
w_jbbgm	Month started current job	UKHLS
w_jbbgy	Year started current job	UKHLS

Note: X can go from 1 to 16. In Wave 13, X is from 1 to 10.

Table 6.3: Variables dropped from the Second Jobs module

Variable	**Variable Label**
w_j2has	Has a second job
w_j2semp	Gross earnings from second jobs last month
w_j2hrs	No. of hours worked per month, second job
w_j2soc90	2nd current job: SOC 1990
w_j2soc90_cc	2nd current job: SOC 1990, condensed
w_j2soc00	2nd current job: SOC 2000
w_j2soc00_cc	2nd current job: SOC 2000, condensed
w_j2soc10	2nd current job: SOC 2010
w_j2soc10_cc	2nd current job: SOC 2010, condensed
w_j2nssec8_dv	2nd job: NSSEC 8 classes

Table 6.4: Variables about Other Jobs in new module ‘Otherwork_w13’

w_owpayu_X	Other work usual pay: Job X
w_owpayug_X	Usual pay in other work: gross/net of deductions: Job X
w_owwah_X	Other work at home: Job X
w_owmaintrv_X	Other work main travel: Job X

Note: X can go from 1 to 16. In Wave 13, X is from 1 to 10.

Table 6.5: New variables in job search module (job number indexed by suffix "_Y")

Variable	Variable Label
w_julksoc90_Y	Job Y searching for: SOC 1990
w_julksoc00_Y	Job Y searching for: SOC 2000
w_julksoc10_Y	Job Y searching for: SOC 2010
w_julksoc20_Y	Job Y searching for: SOC 2020

Note: Y is from 1 to 3.

Table 6.6: Variables in the non-employment module that have not changed

Variable	Variable Label
w_jlsoc00	Last job: SOC 2000
w_jlsoc00_cc	Last job: SOC 2000 (condensed 3 digits version)
w_jlsoc10	Last job: SOC 2010
w_jlsoc10_cc	Lastjob: SOC 2010 (condensed 3 digits version)
w_jlsoc90	Lastjob: SOC 1990
w_jlsoc90_cc	Lastjob: SOC 1990 (condensed 3 digits version)
w_jlsic07	Lastjob: SIC 2007
w_jlsic07_cc	Lastjob: SIC 2007 (condensed 2-digit version)
w_jles2000	Lastjob: Employment Status 2000
w_jlseg_dv	Lastjob: Socio-economic Group
w_jlrgsc_dv	Lastjob: Registrar General's Social Class
w_jlnssec_dv	Lastjob: NS-SEC
w_jlnssec8_dv	Lastjob: Eight Class NS-SEC
w_jlnssec5_dv	Lastjob: Five Class NS-SEC
w_jlnssec3_dv	Lastjob: Three Class NS-SEC
w_jlisco88	Lastjob: International Classification of Occupations 1988
w_jlisco88_cc	Lastjob: ISCO88 (condensed 3 digits version)

Revisions to the main current job occupation (jbsoc00) in Wave 13

From Wave 2 to 12, the **jbsoc00** question about the occupation of the main current job (an open-ended description of the job title and tasks and duties, subsequently coded to the Standard Occupational Classification (SOC)) was preceded by the **jbsoc00chk** question. Note that the **jbsoc00** question was:

> **jbsoc00**: "What was your main job last week? Please tell me the exact job title and describe fully the sort of work you do".

The **jbsoc00chk** question was asked of all respondents who had a job, and whose answer to the **jbsoc00** in the previous wave was codable. In the **jbsoc00chk** question, the respondent was asked if the description of their occupation given previously (the actual text description provided by the respondent was displayed in full on the screen) was still accurate.

> **jbsoc00chk**: Is [ff_JBSOC00] still an accurate description of your occupation in your main job?

If the respondent answered 'yes,' then the **jbsoc00** question was not asked, and the occupation SOC code from the previous interview was copied over (fed-forward). If the respondent said 'no,' (which ranged between 12 – 18 %) then the **jbsoc00** question was asked. Therefore, for around 80 – 90 per cent cases, the previous occupation code was copied over. Using this method of dependent interviewing for the main job occupation question, an occupation change was registered at around 10 per cent of respondents in continued employment. [1]

The way we asked about current employment was changed in Wave 13. Instead of asking about the main job and the second job in two separate modules, respectively, "Current Employment " and "Second Jobs", respondents were asked about all jobs they had (up to 13) in the modified "Current Employment" module (see Changes to employment questions from Wave 13). The **jbsoc00chk** question could not be asked in Wave 13 because, unlike in prior waves [2], we asked a series of **jbsoc00** questions in which the respondent was prompted to describe the occupation of each of the jobs they reported:

> "Please describe fully the sort of work you do for your job as a/an [JOBCODE] {if MULTIJOBSTOTAL > 1}."

Although we could identify which of these was the main job, because of the changes in wording, we could not ensure that prior waves matched Wave 13. Moreover, given that we were asking about all other jobs in the same manner (which had not been done before) we wanted consistent measurement within waves across all jobs as well. Also note that in Wave 13, the **jbsoc00** questions (one for each job reported) only asked to describe the job while the job titles were collected separately in the questions **alljobstitle**. So, for each job reported, respondents were asked:

> **alljobstitle:** "What is the exact job title for your job"

> **jbsoc00:** "Please describe fully the sort of work you do for your job as a/an[jobcode]"**.**

In other words, every employed respondent provided a description of their current main job occupation (the job title, tasks, and duties). These answers were then coded independently of the previous description, or the previous occupation SOC code reported in the previous wave. We found an unusually high percentage of occupation changes in the main job occupation SOC codes among the sample in continued employment (compared to previous waves). There were three reasons for this. First, there was an accidental omission of parts of the available information used in coding the occupations – only the description of the task and duties (**jbsoc00**) was coded, excluding the job title (**alljobstitle**) information. This led to lower overall precision of coding. Second, this procedure was also inconsistent with the pre-wave 13 method, where both the job title and the task description were coded together. Third, the increase in occupation changes was due to the variability inherent to coding answers to open-ended occupation questions.

Two steps were taken to address these issues. First, all **jbsoc00** verbatim responses about the job description were concatenated with the job titles reported in **alljobstitle** and coded using SOC

[1] This percentage could be lower than the share of 'no' answers in the **jbsoc00chk**, as some respondents who answered 'no' provided a description similar enough to the previous one to be coded with the same SOC code as before.

[2] There were two separate questions about occupation, main job: **jbsoc00**, second job: **j2soc00**.

coding frame. This eliminated one source of error and reduced the proportion of job changes observed between earlier waves and Wave 13, from 60 to 45 per cent of the sample in continued employment. Since this was still much higher than the percentage of changes observed pre-wave 13, a further step was taken. For all cases where a change of SOC code was observed, a comparison of the fed-forward occupation verbatim responses (i.e. the title and description of the job collected in the last interview) with the occupation verbatim response collected in wave 13 was undertaken. Here, we aimed to emulate the logic of the pre-wave 13 questionnaire. Although the **jbsoc00chk** question was not asked, meaning there was no respondent's self-assessment of how accurately their last description reflected their current job, we still had the last and the current descriptions. By comparing the two descriptions, we attempted to assess how each respondent would have responded to the **jbsoc00chk** question had they been asked. If the verbatims were similar and described the same occupation, the fed-forward SOC code was assigned. If the verbatims were not similar and thus signalled an occupation change, the new SOC code was assigned. The revision was applied to all previously released versions of the main current job occupation variables: **m_jbsoc00**, **m_jbsoc00_cc**, **m_jbsoc10**, **m_jbsoc10_cc**.

As a result, 70 per cent of the occupation changes were assessed as spurious (the verbatims were similar), with the remaining 30 per cent assessed as genuine (the verbatims were dissimilar). This translated into an overall rate of change of 15 per cent in the sample in continued employment.

Table x. Example (fictitious) answers to the **jbsoc00** question and the comparison result.

Fed-forward jbsoc00 verbatim	Wave 13 jbsoc00 verbatim	Comparison result
a farmer who farms land	farmer, farming land	similar
serving customers in a bar	I work in a nursery looking after children	dissimilar

Derived Income variables

This section introduces the Understanding Society derived income variables and is organised as: Individual income variables; household income variables and household deductions (housing costs and local taxes); and imputation flags and top-coding. A discussion of the use of price indices and equivalence scales is also contained in the "household income variables" subsection. (Other derived variables are discussed in the Derived variables section).

At each wave Understanding Society collects the same detailed information on personal incomes. All individuals aged 16 or over are asked to report receipt of:

- earnings from main and second jobs
- social security benefits
- state and private benefits
- private transfers and investment income

Derived from these reports of the individual income sources, and included in the publicly available datasets, are a number of variables corresponding to different income concepts. The concepts are constructed at either the individual or household level, net or gross of taxes, and with or without household level deductions where appropriate. Examples include: gross individual income, individual net earnings, gross household income and net household income.

Most analysis of living standards, income dynamics, poverty and low income tend to use net or disposable household income after taxes and other deductions i.e., the income that people have available for consumption or saving. Users interested in analysing net household income after taxes should refer to the subsection on household income estimates (and in particular the variable ***w*_fihhmnnet3_dv**). In deriving net incomes, UKHLS follows the approach used by the Department for Work Pensions (DWP) for their Households Below Average Income (HBAI) datasets. UKHLS household income variables are now used as the data source for official UK statistics on income dynamics by the DWP.

This section also provides guidance on how to equivalise household income variables to take account of differences in household size and composition and guidance on how to make household incomes comparable across time using price indices.

Tips for analysts:

It's important to know what income variables are available but also to know how they were constructed. A comprehensive guide to using the income data is discussed in our working paper Understanding Society and its income data which we recommend all income users consult.

Individual Income variables

This section summarises the individual or personal income derived variables. Individual income estimates are included in the individual level data files, ***w*_indresp**.

Total estimated net monthly income is included in the variable ***w*_fimnnet_dv** where "net" refers to net of taxes on earnings and national insurance contributions. It is constructed as the sum of the six income components described below. Gross monthly income, ***w*_fimngrs_dv,** is also estimated from the individual income components described below except that the earnings components are gross, that is, before taxes and National Insurance contributions are deducted and tax is deducted from non-pay income (rental income), which is assumed to be reported gross. The associated imputation flag for both variables is ***w*_fimngrs_if**.

The individual level data files contain estimates of the six components of individual net income. These are as follows:

Component 1: Labour income (***w*_fimnlabnet_dv**)

This is the sum of three earnings components: net usual pay (***w*_paynu_dv**); net self-employment income (***w*_seearnnet_dv**); net pay in second job (***w_j2paynet_dv***).

During the COVID-19 pandemic Government support for employees (furlough) and self-employees (Self-employment Income Support Scheme (SEISS)) are included in the respective pay components. Further details are included in the document Understanding Society changes to the main study due to the COVID-19 pandemic.

Component 2: Miscellaneous income (***w_fimnmisc_dv***)

This includes receipts reported in the income data file where ***w*_ficode** equals [24] “educational grant (not student loan or tuition fee loan)”, [27] “payments from a family member not living here”, or [38] “any other regular payment (not asked in Wave 1)”, [42] “Student Loan and/or Tuition Fee Loan”. This is assumed to be reported net of tax.

Component 3: private benefit income (***w_fimnprben_dv***)

This includes receipts reported in the income data file where ***w*_ficode** equals [25] “trade union / friendly society payment”, [26] “maintenance or alimony”, or [35] “sickness and accident insurance”. This is assumed to be reported net of tax.

Component 4: investment income (***w_fimninvnet_dv***)

This includes receipts reported in income record where ***w*_ficode** equals [4] “a private pension / annuity”, [28] “rent from boarders or lodgers (not family members) living here”, or [29] “rent from any other property”. To this is added the monthly income from savings and investments, estimated as the annual income from savings and investments **(*w*_fiyrinvinc_dv),** divided by 12. All these sources are assumed to be reported net except for rent from other property which is assumed reported gross, and a tax liability is deducted.

Component 5: pension income (***w_fimnpen_dv***)

This includes receipts reported in the income data file where ***w*_ficode** equals [2] “a pension from a previous employer”, or [3] “a pension from a spouse’s previous employer”. This is assumed to be reported net of tax.

Component 6: social benefit income (***w_fimnsben_dv***)

This includes receipts reported in income record where ***w*_ficode** equals [1] “state retirement (old age) pension”, [5] “a widow’s or war widow’s pension”, [6] “a widowed mother’s allowance / widowed parent’s allowance”, [7] “pension credit (includes guarantee credit & saving credit)”, [8] “severe disablement allowance”, [9] “industrial injury disablement allowance”, [10] “disability living allowance”, [11] “attendance allowance”, [12] “carer’s allowance (formerly invalid care allowance)”, [13] “war disablement pension”, [14] “incapacity benefit”, [15] “income support”, [16] “job seeker’s allowance”, [18] “child benefit (including lone-parent child benefit payments)”, [19] “child tax credit”, [20] “working tax credit (includes disabled person’s tax credit)”, [21] “maternity allowance”, [22] “housing benefit”, [23] “council tax benefit”, [30] “foster allowance / guardian allowance”, [31] “rent rebate (NI only)”, [32] “rate rebate (NI only – offset against rates)”, [33] “employment and support allowance”, [34] “return to work credit”, [36] “in-work credit for lone parents”, [37] “other disability related benefit or payment”, [39] “income from any other state benefit (not asked in Wave 1), [40] “universal credit” (from Wave 4), [41] “personal independence payments” (from Wave 4). This is assumed to be reported net of tax.

Personal gross monthly income can be decomposed into three subcomponents: labour income (**w_fimnlabgrs_dv***), equal to the sum of gross usual pay (**w_paygu_dv*)**, self-employment pay (**w_seearngrs_dv*)** and gross second-job pay (**w_j2pay_dv**); annual income from savings and investments (**w_fiyrinvinc_dv**)/12; and monthly income from benefits and other sources (**w_fibenothr_dv**).

Less commonly, a researcher may be interested in the individual components of unearned income, such as individual benefit amounts. These are recorded in a separate data file (**w_income**). This file lists all reported unearned sources. There may be multiple receipts of income from the same source in this file. For example, a respondent may have multiple pensions from a previous employer. Multiple receipts of a given income source are summed. These sums are recorded into the variable **w_frmnthimp_dv**. Therefore, for a given income source, **w_frmnthimp_dv** is equal to the total value of all receipts from that source for the first income receipt, it is set to zero for the subsequent receipts. The same income source may get reported by more than one member of the household. To avoid double counting when constructing household income totals, the derived variable **w_frjtkeep_dv** identifies which one should be included in the household total. Note, the **w_income** file does not include individual income amounts for non-respondents in responding households.

Tips for analysts: COVID-19

* Government support for employees (furlough) and self-employees (Self-employment Income Support Scheme (SEISS)) are included in the respective pay components. Further details are included in the document Understanding Society changes to the main study due to the COVID-19 pandemic.

Household income variables

This section summarises the household income derived variables. Household-level income measures are included in the household level data files, ***w*_hhresp.**

In order to compare incomes for households of different size and composition, each net household income value should be adjusted by an equivalence scale. The public release files contain values of the OECD-modified equivalence scale for each household (**w_ ieqmoecd_dv**). Equivalisation can be performed by dividing each household's income value by the equivalisation value provided.

At the household level, ***w*_fihhmnnet1_dv+** is the net household monthly income. It is the sum of net monthly incomes from all household members (including proxies and within household non-respondents, see ***w*_fimnnet_dv+**). It can be decomposed into the six subcomponents: net labour income (**w_fihhmnlabnet_dv+**), miscellaneous income (**w_fihhmnmisc_dv), ** private benefit income (**w_fihhmnprben_dv),** investment income (**w_fihhmninv_dv),** pension income (**w_fihhmnpen_dv)**, and social benefit income (**w_fihhmnsben_dv**). The six subcomponents map to the six subcomponents of individual net income described in more detail in the Individual Income subsection.

The variable ***w*_fihhmngrs_dv+** is total household gross income. It is the sum of gross monthly incomes from all household members (including proxies and within household non-respondents, see **w_fimngrs_dv**). The variable **w_fihhmnlabgrs_dv+** is gross household labour income.

Income components are imputed for all proxy and within household non-respondents. Hence in Understanding Society household income estimates are available for all households – including where some household members non-respond. Users may decide to drop cases based on such imputed data but they would then need to adjust their results to take into account the consequent sample selection i.e. their results would no longer be representative of the UK population. Details of how to identify imputed cases are provided in the imputation flags section.

Tips for analysts: COVID-19

+Government support for employees (furlough) and self-employees (Self-employment Income Support Scheme (SEISS)) are included in the respective household pay components. Further details are included in the document Understanding Society changes to the main study due to the COVID-19 pandemic.

Housing costs (Under review in 2023)

In the UK a distinction is sometimes made between incomes before and after housing costs, where housing costs usually include rent, mortgage payments, structural insurance premiums and local water charges. To allow for the computation of income after housing cost, the Understanding Society has a set of such variables that focus on rent and mortgage payments.

The final housing costs derived variables are as follows: For renters, **w_rentgrs_dv** is the computed monthly gross rent i.e., including any housing benefit received. **w_rent_dv** is the monthly rent net of housing benefit (equal to **w_rentgrs_dv** where no housing benefit is received). Missing values are imputed (see below), and, where the household reports 100% housing benefit (in which case no rent value is reported), the value is set equal to housing benefit reported in the individual questionnaire and a value imputed if not reported there. The variable **w_rentg_if** is an imputation flag for **w_rentgrs_dv**.

In some cases, a reporting inconsistency arises where housing benefit is implicitly reported in the household questionnaire (as the difference between gross and net rent), and it differs from what is reported in the individual questionnaires. The main income variables of the previous sections take the individual questionnaire reports to be correct and so they do not include implicitly reported housing benefit from the household questionnaire. When working with housing costs variables, as they imply an amount of housing benefit, it is necessary to adjust the household income totals to include it. So that data users can adjust the household income totals, an adjustment factor (***w*_hbadjust_dv**) is included in the public release files. For a given household with implied housing benefit in the household questionnaire, this factor is equal to housing benefit reported in the household questionnaire minus the sum of housing benefit reported in the individual questionnaires. Household income totals can therefore be adjusted by adding the adjustment factor to the total household income. Gross household income has already been adjusted in the variable **w_fihhmngrs1_dv** (=**w_fihhmngrs_dv** + ***w*_hbadjust_dv**).

For those paying mortgages, ***w*_xpmg_dv** is monthly total mortgage payments including imputation. The variable ***w*_xpmg_if** is the imputation flag for this variable. Most definitions of housing costs for purposes of measuring income after housing costs seek to exclude repayments of capital included in mortgage payments and only include interest payments. ***w*_xpmgint_dv** is the estimated interest within ***w*_xpmg_dv**. For short period mortgages it is based on data on current interest rates times the outstanding principal and for mortgages with more than two years to run based on a standard repayment mortgage formula.

The above variables for rent and mortgages are combined in the following variables: **w_houscost1_dv** is total housing costs including capital repayments i.e., **w_rentgrs_dv** + **w_xpmg_dv**. **w_houscost2_dv** excludes capital repayments, i.e., **w_rentgrs_dv** + **w_xpmgint_dv**.

The imputation of rent and mortgage payment assumes that variations over time are small and where other reports at the same address are available in other waves, missing values are set equal to the median of these reports. Where no report at that address is available, a single value is imputed based on characteristics of the accommodation and household (including region, number of bedrooms, gross income, household composition and age, rural/urban indicator) and applied to all relevant waves

Local taxes

Council tax is a UK local tax that is often deducted from gross household income in poverty analysis. In Understanding Society, council tax liability is currently estimated for Great Britain, though not for Northern Ireland. Council tax liability (net council tax) is included in the variable **w_ficountax_dv**. It is equal to gross council tax applying any single person discount and then subtracting any council tax reduction. The variable ***w*_fihhmnnet3_dv** (only available in *Special License data*) is equal to ***w*_fihhmnnet1_dv** less council tax liability and any council tax reduction (not released) . ***w*_fihhmnnet4_dv+** (only available in *Special License data*) is equal to ***w*_fihhmnnet3_dv+** and adjusted for housing benefit reported in the household questionnaire.

Price indices

In order to compare household incomes in different months and years, each household income value should also be adjusted by a price index to a common time period. In the publicly released files, none of the Understanding Society income variables have been adjusted to account for price changes over time.

Many price indices are publicly available including the Consumer Price Index (CPI) and Retail Price Index (RPI) from the Office for National Statistics. The price index used in official income statistics –

the Consumer Price Indices series excluding rents, maintenance repairs and water charges - is freely available for download[1] (Tab 1a is the “after housing costs” series i.e. CPI excluding rents, maintenance repairs and water charges. Tab 1b is the “before housing costs” series i.e. CPI excluding MIPs, ground rent and dwellings insurance.)

The value of a price index can easily be merged to a household’s month of interview (**w_intdatey**, **w_intdatem**) in each survey year, so that necessary adjustments to incomes can be performed.

Imputation flags and top-coding of income variables

Imputation flags help you identify derived income values that have been imputed. Imputation is used where a respondent is either missing an individual interview (unit nonresponse) or does not provide an answer to a particular income question (item nonresponse). Exact details of the imputation methods used in Understanding Society can be found in "Understanding Society and its income data".

For each income variable for which amounts are imputed there is a separate imputation flag variable with a suffix “**_if**” instead of “**_dv**” indicating whether the variable is imputed. In most cases the flags take the value 1 if imputed and 0 if not, but in the case of the following variables it shows the proportion of total income imputed: **w_fimngrs_if**, **w_fibenothr_if**, and **w_fihhmngrs_if**.

Top coding of income and investment variables

Because extremely high incomes are rare, releasing this information can increase the risk of disclosure. So, reported and imputed income and investment amounts have been top-coded in order to prevent disclosure of respondents. Individual earnings and self-employment profit as well as investment income and rent have been top-coded at £100,000 per annum or its monthly equivalent less tax and national insurance for net incomes.

Derived total personal income and household income are computed using the top-coded values and there is a set of flag variables indicating whether the computed sum variables include top coded amounts.

The flag variables on the **w_indresp** data file are:

w_fimngrs_tc, **w_fimnlabgrs_tc**, **w_fimnlabnet_tc**, **w_fiyrinvinc_tc**, **w_fibenothr_tc**

The flag variables on the **w_hhresp** data file are:

w_fihhmngrs_tc, **w_fihhmnlabgrs_tc**

The **w_income** data file includes the flag **w_frmnth_tc**.

The following are top-coded at +/- £8,333 per month or its net equivalent:

w_payg_dv, **w_payn_dv**, **w_payu_dv**, **w_paygu_dv**, **w_paynu_dv**, **w_j2pay_dv**, **w_j2paynet_dv**, **w_seearnnet_dv**, **w_seearngrs_dv**, **w_seearnnet_dv**, **w_frmnthimp_dv**, **w_jspayu**, **w_j2pay**, **w_paygl**, **w_payu**, **w_paynl**

The variable **w_fiyrinvinc_dv** is top coded at £100,000 per annum.

Data from the investment income module in Wave 4 (**d_nvestamtrt1**, **d_nvestamtrt2**, **d_nvestamtrt3**, **d_nvestamtrt97**) have been top-coded at £1,000,000.

Access to income and investment data without top-coding is available in the Special License version of the data study number 6931.

Weighting guidance

Understanding Society has a complex design. The dataset allows a vast array of research across different disciplines, topics and population sub-groups.

This section of the user guide provides advice on the correct weight to use and the sample design variables (for clustering and stratification) provided with the data which will allow you to estimate standard errors correctly.

This video provides an introduction to choosing and using weights in Understanding Society. Note this video does not mention the weights for the latest sample added, GPS2.

You can also refer to the weighting guidance within this section as well as search our User Support Forum for questions which have been asked before. If your question hasn't already been answered, you can post one.

Why use weights?

The UKHLS dataset is designed to be used with weights.

The weights adjust for unequal selection probabilities, differential nonresponse, and potential sampling error. Unweighted analysis does not correctly reflect the population structure as some groups may be over-represented in the sample by design (e.g., over sampling of ethnic minorities in the EMBS) or due to nonresponse as certain types of people are more likely to respond than others.

If a variable or characteristic differs by these groups which are over/under-represented in the sample, then estimates of that variable based on this sample will be biased. For example, if the wages of some ethnic minority groups are lower in the population than the white British population and as there is a higher proportion of ethnic minorities in this sample than in the population, unweighted estimates of UK wages will be downwardly biased. A weighted analysis will adjust for the higher sampling fraction in Northern Ireland and for different probabilities of selection in the EMB and IEMB samples, as well as for response rate differences between subgroups of the sample. An unweighted analysis does not correctly reflect the population structure unless the assumptions below are true. It is suggested that researchers publishing or presenting unweighted estimates make these assumptions explicit.

What happens if I don't use a weight?

You implicitly assume that sample members have equal probabilities of selection and of response. This is not true.

If no weighting is used, an analysis of Understanding Society data assumes that all estimated parameters (means, measure of dispersion, model coefficients, etc.) do not differ between:

- Northern Ireland and the rest of the UK
- People of ethnic minority origin and people of white British origin
- Recent immigrants to UK and people who have stayed in the country longer
- People who live at an address with more than three dwellings or more than three households and those who don't
- People who responded at Wave 1 and those who did not

- People who continued to respond at later waves and those who did not
- People who responded to each particular instrument used in the analysis (individual interview, self-completion questionnaire etc.) and those who did not, see Lynn, Burton et al. (2012)

An unweighted analysis of the former-BHPS sample assumes that estimated parameters do not differ between:

- Each of England, Scotland, Wales and Northern Ireland
- People who live at an address with more than three dwellings or more than three households and those who don't
- People who responded at Wave 2 of Understanding Society in 2010 and those who may have become non-respondents at any time since Wave 1 of BHPS in 1991
- People who keep responding in later waves of Understanding Society and those who stopped responding at any point of time between 1991 and the last year in your analysis

We therefore strongly suggest conducting weighted analyses of the Understanding Society data.

Weights are constructed by combining (i) design weights which adjust for unequal selection or sampling fraction and (ii) nonresponse correction which adjusts for differential nonresponse and attrition at various stages (household level, within household at individual level, whether adult respondent completed self-completion questionnaire or not). Nonresponse occurs when some people/households respond while others do not. Differential response occurs when response is related to a particular characteristic/variable of interest.

Selecting the correct weight for your analysis

The UKHLS dataset is designed to be used with weights (To see why visit the Why use weights page).

Separate sets of weights are provided for:

- The combined GPS and EMBS (from Wave 1)
- The former BHPS sample (from 1991 and 2001)
- The combined GPS, EMBS and BHPS (from Wave 2)
- The combined GPS, EMBS, BHPS and IEMBS (from Wave 6)
- The combined GPS, EMBS, BHPS, IEMBS and GPS2 (from Wave 14)

The available sets of weights are not identical for these four analysis bases, reflecting differences in data collection. For any proposed analysis, weights should be selected carefully. Given the complexity and multi-purpose nature of the Understanding Society design we provide multiple sets of weights to meet the different needs of users. The weight for your analysis reflects the survey instrument that is the source of the data being used in the analysis (e.g. household grid, household questionnaire, individual questionnaire, etc.), the analysis level (household or individual), and the combination of waves involved.

Each weight except for design weights has been scaled to have a mean of one amongst cases eligible to receive the weight.

To get started watch our short video about selecting weights in Understanding Society (note this video does not mention the weights for the latest sample added, GPS2). The naming conventions for weights are intended to help users to identify the correct weight. The name of each weight reflects the wave for which the weight is calculated, level of analysis, data source and its nature (design weight, cross-sectional analysis weight or longitudinal analysis weight). There are a number of weights reflecting the complex structure of the data and the help within this section gives guidance on which weights to use. All weight names follow the same structure **w_xxxyyzz_aa**:

The rules are described below.

Naming convention for Understanding Society weights

Wave letter	Who are you studying?	Which questions(naire)?	Which sample/timeline?	Analysing one wave or across waves?
w_ (a to n)	**Xxx** (Hhold or individual)	**Yy** (instrument)	**Zz** (samples cover different waves)	**_aa** (cross-sectional/longitudinal)
a_	**hhd:** household	**en:** enumeration	**us:** GPS & EMB (W1>)	**_xw:** cross-sectional analysis weight
b_	**psn:** persons 0+	**in:** interview	**bh:** BHPS (W2>)	**_lw:** longitudinal weight_
c_	**ind:** persons 16+	**px:** interview or proxy	**ub:** GPS, EMB & BHPS (W2-W5)	**_xd:** x-sectional design weight
d_	**yth:** persons 10-15	**5m:** "extra 5 minutes"	**ui:** GPS, EMB, BHPS & IEMB (W6>)	**_li:** longitudinal inclusion weight
e_		**sc:** self-completion		
f_			**g2:** BHPS, GPS, EMB, IEMB and GPS2 (W14>)	
g_			**91:** BHPS original sample (91> excl. N.I.)	
			01: BHPS original sample + boosts	

Cross-sectional or longitudinal analysis

If your analysis uses only data from one Wave, select the "**xw**" (cross-sectional) version of the weight. This weight is defined for all sample members who responded to the relevant questionnaire at a particular Wave. If your analysis uses data from multiple or consecutive waves select an appropriate "**lw**" (longitudinal) version of the weight.

Example - if your analysis only uses data from Wave 4, select the "xw" (cross-sectional) version of the weight (note all Wave 4 variables begin with **d_**). If your analysis uses data from multiple waves select an appropriate "lw" (longitudinal) version of the weight from the last wave. For example, if you are looking at Waves 4 to 9, use the appropriate longitudinal weight from the last wave in your analysis (note Wave 9 variables begin with **i_**). The longitudinal and cross-sectional weights page explains the difference between the weights and when to use them.

Hierarchy of analysis levels

For individual level analysis you may want to combine information from different questionnaires. In this situation please select the weight suitable for the lowest level according to the hierarchy table below:

Hierarchy table

Levels of analysis	Data source	_xxxyy
5	Household grid and/or household questionnaire	_psnen
4	Adult proxy and main interview	_indpx
3	Adult main interview only (no proxy)	_indin
2	Adult self-completion interview	_indsc
2	Extra 5 minutes interview	_ind5m
2	Youth questionnaire	_ythsc

Example - if you are analysing cross-sectional data from Wave 1, and using questions from both the proxy/full interview as well as from the self-completion questionnaire, then the correct weight will be **a_indscus_xw** – the weight for the self-completion questionnaire is level 1 in the table and is lower than the proxy/full interview questionnaire at level 3.

Variable weight **a_indscus_xw** is designed for participants from Wave 1 **(a_)**, aged 16+ **(ind)**, who answered via the self-completion questionnaire **(sc)** from the general population and ethnic minority samples **(us)** and analysed within one wave **(xw)**.

After you have decided the population you want to generalise your results to and the questionnaire(s) you want to use, refer to these tables to decide the appropriate weight for your analyses.

Alternatively, you can refer to the Naming conventions table above that summarises the naming convention and makes it easy to choose the correct weight

For advanced users who want to model nonresponse in their own way, we provide design weights and inclusion weights which adjust the sample for unequal selection probabilities but not for nonresponse. Note that adjusting for the first wave nonresponse is different from adjusting for attrition and requires variables which have values for both responding households and never responding households.

Longitudinal and cross-sectional weights

Two main types of weights are released with UKHLS data: longitudinal weights and cross-sectional weights.

Longitudinal weights

Longitudinal weights should be used for any analysis that includes information from more than one wave. Longitudinal weights are created for monotone longitudinal response, meaning that households or respondents (depending on the weight) participate in each previous wave. This also means that for nonmonotone longitudinal analysis, where some panel members have missed one or more waves previously but have participated in other waves that are used in a particular analysis, the weight provided will give representative estimates but will not use all the available information (and therefore may have potentially lower statistical power than if a tailored weight is created for such analysis). In most such situations the sample size is still sufficient for analysis with the longitudinal weights provided, but where there may be concern about statistical power, a tailored longitudinal weight may be created by the user to reflect their specific combination of waves. For this you can follow our online course on Creating Tailored Weights. Note, that analysis with a tailored weight by expectation should give the same point estimates as analysis with the provided longitudinal weights but should result in slightly narrower confidence interval around these estimates.

Cross-sectional weights

Unlike in a cross-sectional study where participants are selected at the time of interest, in a longitudinal study participants are selected at one point of time, though their information can be used cross-sectionally years later. Cross-sectional weights in a longitudinal study allow a user to represent a population cross-sectionally with some exceptions (note that a longitudinal study does not include people who spend a few months, or even a few years in the country, specifically recent immigrants since the last boost and their children are excluded from the Study, and if they remain in the country this is true until the following boost). Cross-sectional weights should only be used if all the information in an analysis comes from one wave.

Cross-sectional weights are created based on longitudinal enumeration weights. Enumeration represents response at a household level and gives a weight to all household members even if only one person in the household responds in a particular wave. Longitudinal enumeration weights are only available for OSMs (original sample members, those that were selected into the Study or their children born into the Study). But given that UKHLS is a household study and interviews all household members, it therefore also gathers information on TSMs (temporary sample members, those that were not in a household at the beginning, but joined a household sometime during the course of the Study). Information from TSMs is incorporated into the information from OSMs through a weight-share method. Any other cross-sectional weight is then modelled from the enumeration weight. Information on the sample design for Understanding Society can be found in the Study Design section.

Sharing information from OSMs to TSMs results in the situation where everyone who has a valid answer in a particular wave and is part of a continuously enumerated household has a positive non-zero cross-sectional weight. Yet there will still be some households with some missing waves, even though in a current wave of interest some or all household members have completed interviews. With each year a chance of a household missing one or more waves in the past but still participating is increasing, which results in an increase in potential zero cross-sectional weights. This situation is resolved through another stage of weight-share, although importantly there is a set of panel members whose weights always remain zero. This group is called TSMs from Wave 1, and they represent people who were not eligible for selection into Ethnic Minority Boost at Wave 1 of UKHLS or into Immigrant and Ethnic Minority Boost at Wave 6 of UKHLS, but who live in the same household with people eligible for selection into these boosts. TSMs from Wave 1 are different to those people who join household later in the Study (TSMs) as their selection probabilities are set to 0 from the start, and this defines their later weights which always have to be zero. Interviewing these people is still important as their information provides context to understanding immigrants and ethnic minority household life. Yet their interview information cannot be used in a similar way to that of other panel participants, which results in their zero cross-sectional weights. Note, that the number of TSMs from Wave 1 is very small and is decreasing with time as when such people leave the household they are not followed.

Representing the population

Understanding Society can be used in different ways, to represent several different populations.

You can represent the cross-sectional population (those currently resident in the country) in any year since 1991 or the longitudinal population over a series of years (those continuously resident in the country over a period of time). You need to identify the appropriate data files and the appropriate weight to use, depending on the population you wish to represent. There are some important points to note: From 1991 to 2000, the Study only covered Great Britain (England, Scotland and Wales). It was extended to Northern Ireland in 2001. Consequently, you can represent:

- the cross-sectional population of Great Britain in any year since 1991
- the longitudinal population of Great Britain over any period of years since 1991
- the cross-sectional population of the United Kingdom in any year since 2001
- the longitudinal population of the United Kingdom over any period of years since 2001.

However, a much larger sample size is available from 2009-10 onwards, when data collection from the main Understanding Society samples (General Population Sample and Ethnic Minority Boost Sample) started, so longitudinal analysis starting at this point can be particularly valuable for the study of small subgroups or rare events. Due to the sampling methods used, some recent immigrants are excluded from several of the possible reference populations. The only populations with no such under coverage are Great Britain in 1991, Wales and Scotland in 1999, Northern Ireland in 2000, UK in 2009-10 and in 2014-15: The data collected between 1992 and 2008 in England exclude households consisting entirely of recent (since 1991) immigrants. In Wales and Scotland, data collected between 1992 and 1998 exclude households consisting entirely of immigrants since 1991 and data collected between 2000 and 2008 exclude households consisting entirely of immigrants since 1999. In all countries of the UK, data collected between 2010/11 and 2013/14 exclude households consisting entirely of immigrants since 2009/10, and data collected since 2015/16 exclude households consisting entirely of immigrants since 2014/15.

Representing a subpopulation

You can represent any subpopulation of any of the populations described above, provided it is defined by substantive variables. If you use appropriate analysis methods for the relevant population, but restrict your analysis to members of the subpopulation, your results will be representative of the subpopulation.

Examples of subpopulations that you can represent:

- Residents of Northern Ireland
- Females in full time employment
- Babies born in the last 12 months
- Conservative voters in the 2017 election
- Males aged between 17 and 29 who hold a driving license

Are sample sizes adequate to represent ethnic minorities or immigrants?

In 2009-10 (Wave 1 of Understanding Society) data were collected for the first time from an ethnic minority boost sample which was designed to provide substantially boosted sample sizes for the following subgroups: Indian, Pakistani, Bangladeshi, Afro Caribbean and Black African. From 2014-15 (Wave 6) we further boosted the five ethnic minority groups listed above and also added a boost of immigrants (i.e. persons born outside of the UK). If you are interested in immigrants other than of the five ethnic groups listed above you may want to start your analysis from Wave 6. These subgroups are also asked additional questions, referred to as the "extra 5 minutes" questionnaire. These additional questions are also asked of a small random 4 subsample of the general population sample which can be used to compare findings for ethnic minority groups to the total population. For this use the weight ***w*_ind5mus_*aa***.

How to represent a population or subpopulation

To produce population (or sub-population) estimates you need to apply weights and compute estimate standard errors taking into account the complex survey design. The Analysis guidance for weights, PSU, Strata page shows how to do this.

Analysing a subset of a population – what weights should I use?

It depends whether the subset is defined by personal characteristics (e.g. sociodemographic or geography).

It is appropriate to use the provided weight, even though this was derived for the whole sample. Though not tailored specifically to your analysis sample, the provided weights should not only make the total sample representative of the total population but should also make any subset of the sample representative of the equivalent subset of the population. For example, sample members resident in Wales will represent the population of Wales; female sample member born between 1951 and 2000 will represent all women in the population born between 1951 and 2000, and so on.

In the case of using an unusual combination of instruments for which we do not provide a weight, you have a choice between three options:

- Use the weight provided for the (smallest) hierarchically-superior (larger) sample
- Use the weight provided for the (largest) hierarchically-inferior (smaller) sample

- Derive your own weight, tailored to your analysis sample (take a look at our Open Essex (MoodleX) course Creating tailored weights for UKHLS).

The first two options are both sub-optimal, in different ways, but are simple to implement and the sub-optimality may be minimal. With the first option, the weights will be correcting for a different nonresponse process to the one relevant to your analysis sample.

With the second option, weights will not be defined for all potential members of your analysis sample, but the weights will correct for the relevant nonresponse process. In the example, the largest hierarchically-inferior sample for which weights are provided is the set of people who gave a full interview, including the self-completion component, at all of Waves 1 to 5, **e_indscus_lw**. If this weight were only defined, for example 15,613 out of the 17,977 members of your potential analysis sample, i.e. 86.8%, using this weight would cause a (very slight) loss of precision (but will not introduce additional bias).

Deriving your own tailored weights

If your analysis sample is a nonresponse-defined subset of a sample for which analysis weights have not been provided, you can derive your own analysis weight.

When to derive your own weight

You may consider doing this if no analysis weight has been provided for the combination of waves and instruments that you wish to include in your analysis and if the solutions suggested in this guide are not satisfactory.

For example, suppose you wish to carry out longitudinal analysis of responses to questions that were included at Waves 1, 4 and 7. Your analysis base is therefore sample members who completed an individual interview at each of those three waves (let’s assume that your survey questions of interest were not all included in the proxy questionnaire and that you therefore cannot include proxy responses in your analysis).

One option would be to use the Wave 7 longitudinal weight for the wave 1 sample, i.e **.** **g_indinus_lw**. However, this weight is only defined for sample members who gave a full personal interview at all seven waves, thus 18,510 persons have this weight, whereas 20,390 responded at Waves 1, 4 and 7 (so, 1,880 of those who responded at Waves 1, 4 and 7 must have failed to respond at one of Waves 2, 3, 5 or 6). Using this weight for your analysis would therefore cause almost 10% of your potential analysis sample to be dropped from the analysis. This reduction in sample size will cause a modest reduction in the precision of your analysis (increase in standard errors). The effect will be rather small, and you may well be willing to accept this slight reduction in sample size, unless you are producing estimates for very small population subgroups. But if you want to be able to include all 20,390 respondents in your analysis, you will need to derive your own weight.

How to derive your own weight

First, identify the (smallest) hierarchically-superior sample for which weights have been provided. In this example case, it is the wave 1 responding sample. For this sample, the weight **a_indinus_xw** has been provided. This will serve as your "base weight", to which you will make an adjustment tailored to your analysis sample.

Next, fit a conditional weighted model (e.g. logit) of response to your wave-combination of interest. In the example case, the base for the model would be all Wave 1 responding OSMs (i.e. OSMs with a non-zero value of **a_indinus_xw**) and the dependent variable would be a 0/1 indicator of whether they also responded at both Wave 4 and Wave 7 (and removing from the base any known to have died or emigrated before Wave 7). Predictor variables in the model can be anything relevant observed at wave 1. We also suggest you correct for mortality as described in the Creating tailored weights for UKHLS course. The model will give you a predicted probability for every wave 1 respondent of responding also at waves 4 and 7. Call this Pi.

To make the adjustment to your base weight you multiply **a_indinus_xw** by 1/Pi for all the cases in your analysis sample.

If you are creating weights for the first time this online training course provides guidance on the process: Creating tailored weights for UKHLS.

Clustering and stratification

As the sample design involves stratification and clustering, these design features affect standard errors and should therefore be taken into account in analysis. Appropriate variables are provided to allow the analyst to do this. Here we describe the stratification and clustering variables in the main Understanding Society data files. See Fumagalli, Knies et al. (2017) for a description of these variables in the harmonised BHPS. General advice on using this information appropriately applies to both the UKHLS and harmonised BHPS data.

The variable indicating the primary sampling unit is **psu.** It is available in the cross-wave files, **xwavedat xwaveid**. As the PSU is determined at the time of sampling the value of this variable does not change over time. But to make it easier to use, this variable is also included in wave specific data files, where the name of the variable is ***w*_psu** with "*w*_" reflecting the wave prefix. Similarly, as stratification occurs at the sampling stage, the variable representing stratification, **strata**, does not change over time and is available in the cross-wave files, **xwavedat xwaveid**, and in wave specific data files with a wave prefix, **w_strata.**

Description of variables

These tables provide details of the range of values for variables w_psu and w_strata.

What happens if I don't correct for clustering?

Taking sample clustering into account is simple to do in most standard statistical software for most kinds of estimation. However, if you do not do this, while your estimates are not affected, associated standard errors will tend to be underestimated – sometimes considerably so – resulting in biased hypothesis tests and overfitting of models (with the exception of the distribution of gender where the effect of clustering has sometimes been shown to decrease confidence interval, but such effect is usually observed where households comprise highest level of clustering).

What happens if I don't correct for stratified sampling?

Taking the stratified nature of the sample design into account is simple to do in most standard statistical software for most kinds of estimation. However, if you do not do this, your estimates are not affected, but associated standard errors will tend to be slightly over-estimated. This makes your analysis slightly conservative, which is often acceptable.

Analysis guidance for weights, psu, strata

As discussed in other sections, Understanding Society is a probability survey with a complex sample design and most of the sub-samples were clustered and stratified with unequal selection probabilities (i.e., not all population sub-groups are selected with the same probability). Most statistical softwares assume that the data is from a survey where the sample design is SRS and all sub-groups are selected with equal selection probability and random attrition and nonresponse. So, estimates and their standard errors produced using Understanding Society data without any further adjustments may be biased.

The estimates will be biased in favour of groups who are over-represented in the sample (compared to the population) if the variable statistic being estimated differs by this group. For example, as average pay is lower for most ethnic minority groups as compared to the White majority group, and the former are over-represented in the sample (due to EMBS and IEMBS), then the UK average pay estimated using this data will be underestimated. But as the weights provided are designed to counteract this, weighted estimates will be unbiased estimates of population statistics.

Standard errors of estimates produced from a sample with a clustered design is likely to be *higher* than that of estimates produced from a sample with a SRS design of the same size. The opposite is the case for stratified samples. As most statistical softwares assumed SRS design, without further adjustments the estimated standard errors of estimates will be incorrect.

Most statistical softwares have specific commands that allow you to specify these features. In the case of Stata it is the *SVY* suite of commands, in SPSS it is the *Complex Samples* suite of commands, in R it is the *Survey* package and in SAS it is *surveymeans* command. Please take a look at the section on 'Working with weights and complex survey design' in our online courses 'Introduction to Understanding Society using Stata, SPSS, SAS and R'. There is a different Moodle course for each software, so choose the one based on the software you use. In this section, we provide a worksheet with a worked out example to help you understand how to produce weighted estimates with correct standard errors using that software. The accompanying syntax and output files are also provided. For example, to produce unbiased estimate of average monthly pay in the UK in 2009-10, with correct standard errors using our data with Stata, you will need to do the following:

```
use a_indresp, clear
```

```
svyset a_psu [pweight=a_indinus_xw], strata(a_strata)singleunit(scaled)
replace a_paygu_dv=. if a_paygu_dv<0
svy: mean a_paygu_dv
```

Each Understanding Society weight is set to zero for all sample units to which it does not apply. Thus, specifying the use of the correct weight in analysis will automatically result in the analysis being restricted to the appropriate sample. For example, there are around 2,000 persons in the file **h_indresp**, with a zero value of **h_indinui_lw**. The persons with non-zero values of this weight variable are the people who gave a full individual interview at all of Waves 6, 7 and 8 and the waves before these. For longitudinal analysis of data obtained in the individual interviews at Waves 6, 7 and 8, it is therefore sufficient to specify use of the weight **h_indinui_lw** or to create your own tailored weight. The analysis sample can of course also be further restricted by selecting based on respondent characteristics (e.g. by gender, age, ethnicity, employment status etc): the weight is appropriate for analysis of any demographic subset of the full sample to which the weight applies. Please see 'Selecting the correct weights' section to know about all the different weights that are provided and how to select the correct weight for each type of analysis.

The weights provided have been developed for use when analysing data from various combinations of survey instruments in one of two ways:

When using data from a series of consecutive waves, e.g. a panel analysis. These are the longitudinal weights ending in **_lw**;

When using data from a single wave. These are the cross-sectional weights ending in **_xw**.

Reference

West, B.T,, Sakshaug, J.W. and Aurelien, G.A. (2018) Accounting for complex sampling in survey estimation: a review of current software tools, Journal of Official Statistics 34(3): 721-752. doi: 10.2478/JOS-2018-0034.

Analysis guidance for weights when fitting multilevel models

There are no optimal solutions to fitting multilevel models to survey data, but some work better than others. Two-level models where the higher level corresponds to clusters in the sample design are the only models supported by developed theory. Other models can only be based on intuition and we recommend caution.

Suppose you are fitting a two-level linear model with individuals at level 1 and level 2 corresponds to household. Households correspond to the units in the penultimate sampling stage in the general population.

For some users, it could be tempting at this stage to appeal to the notion that one does not need to worry about the survey weights because clustering – an important aspect of the sampling design – has already been accounted for in the model. However, this line of argument is not recommended because other aspects of the design (e.g., stratification, clustering in the primary sampling units (PSU) at the postcode sector level, variation in selection probabilities) have not been included.

Generally, unless you are sure that **every** aspect of the complex sampling design has been included in the model then the survey weights should always be used.

Therefore, we recommend you fit this model using 'pseudolikelihood' estimation (Pfeffermann et al. 1998; Rabe-Hesketh and Skrondal 2006). Pseudolikelihood combines information about the complex sampling design of Understanding Society (UKHLS) and the modelling assumptions implicit in the two-level model to ensure the correct estimates of the model parameters and the standard errors of these parameters are reported.

There is, however, a complication for the pseudolikelihood estimation of two-level models when the analyst is interested in estimates of the variance components (e.g., the variance of the random effects). The overall survey weight must be split into its two components: Level 1 weight $w_{i|h}$ and Level 2 weight w_h where i represents each individual in cluster (household) h. As the overall survey weight is $w_{ih} = w_h \times w_{i|h}$, the user needs to know either,

(i) Level 2 weight w_h and level 1 weight $w_{i|h}$ or
(ii) Level 2 weight w_h and overall weight w_{ih}.

UKHLS provides data users with (ii) as set out in the example (see Box 1).[3]

If you are using Stata, the easiest way to fit your model using pseudolikelihood is to use the **`mixed`** command. This must be done using the maximum likelihood **not** restricted maximum likelihood (REML) option because pseudolikelihood does not work for REML (at least as it is implemented here).

You do not need to set up the survey function using `svyset`. Instead, you set the following options when using `mixed`:

a) Your Level 1 weight in `[pw = <name of level 1 weight var here>]`
b) Your Level 2 weight in `pweight(<name of level 2 weight var here>)`
c) Make your standard errors robust to PSU clustering by setting `vce(cluster <PSU variable here>)`

As noted above, make sure the `mle` fitting option is specified and not `reml`.

It is recommended that you conduct your analysis using **scaled weights** as set by the `pwscale()` option. Estimating the model without this option runs the risk of producing biased estimates of the variance components (e.g., the random effect variances).

Three scaling options are available each of which promises to reduce bias most effectively in different theoretical scenarios. These options are as follows:

[3] The two types are closely related: type (i) weights can be created from type (ii) weights simply by dividing the overall weight by the level 2 weight for every individual. If you have type (ii) weight, it is recommended you derive and use type (i) weights. In the case of UKHLS, $w_{i|h}$ should be derived as the released individual weight divided by the released household weight.

1. `pwscale(size)` – scales the level 1 weights to equal the sample size.
2. `pwscale(effective)` – scale the level 1 weights to equal the *effective* sample size (that is, the size the sample would be if it were drawn using simple random sampling)
3. `pwscale(gk)` – scales the level 1 weights to be equal and the level 2 weight by the mean level 1 weight.

Although we draw a distinction between type (i) and type (ii) weights, this makes no difference to either option 1 or option 2, which will return identical results. However, `pwscale(gk)` requires the type (i) weights to produce the correct results.

We advise estimating the model using all three scaling options to assess the robustness of your variance components estimates to different choices. All (including the unscaled estimates) should perform similarly (but not identically) in terms of the regression coefficients. In terms of the variance components, the theoretical justification for 1-2 relies on the population size in each Level 2 unit being large, but both have been shown to perform well even when this does not hold; conversely, scaling 3 does not make this assumption but is based on an approximation. Korn and Graubard (2003, table 1) show all three sets of weights are shown to have comparable performance for small population clusters.

Finally, weights are not the whole story. If Level 2 of the model does not correspond to the postcode sector PSUs used by UKHLS, cluster-robust variance estimation must be used to account for clustering (see point c) above). However, it should be noted that the effects of stratification are not taken into account so inferences may be slightly conservative.

For MLwiN users, teaching materials about how to use survey weights are available from the Centre for Multilevel Modelling web site (Pillinger 2011). For three-level models and more generally, the weight scaling is more complicated (e.g., Rabe-Hesketh and Skrondal 2006). MLwiN can be run from Stata.

Box 1
Example using Understanding Society data

Assume PSU as level 1, household as level 2 and individual as level 3.

```
global ms “where UKHLS data is stored”

use "$ms/a_indresp", clear
isvar pidp a_hidp a_age_dv a_jbstat a_mastat_dv a_nchild_dv a_scghq1_dv
a_indscus_xw
keep `r(varlist)'
merge m:1 a_hidp using "$ms/a_hhresp", keepus(a_hhdenus_xw) nogen keep(3)
merge 1:1 pidp using "$ms/xwavedat", keepus(psu strata sex_dv ethn_dv
psnenus_xd) nogen keep(3)

mvdecode _all, mv(-21/-1)
recode sex_dv 0=.

global fe_eq1 i.sex_dv c.a_age_dv##c.a_age_dv i.ethn_dv i.a_jbstat
i.a_mastat c.a_nchild_dv

drop if a_hhdenus_xw==0
drop if a_indscus_xw==0

generat wgtlevel1=psnenus_xd
generat wgtlevel2= a_indscus_xw/wgtlevel1

mixed a_scghq1_dv $fe_eq1 [pw=wgtlevel2] || a_hidp:, mle
pweight(wgtlevel1) vce(cluster a_psu) pwscale(gk)
```

References

Korn E, Graubard G. (2003). Estimating variance components by using survey data. J. R Statist. Soc. B **65** 175-190

Pfeffermann D, Skinner C, Holmes D, Goldstein H, Rasbash J. (1998). Weighting for unequal selection probabilities in multilevel models. J. R. Statist. Soc. B **60** 23-40.

Pillinger R. (2011) Weighting in MLwiN. Centre for Multilevel Modelling Learning Materials https://www.bristol.ac.uk/cmm/team/pillinger-learning-mats.html

Rabe-Hesketh S, Skrondal A. (2006). Multilevel modelling of complex survey data. J. R. Statist. Soc. A **169** 805-827

Analysis advice for mixed mode data

Using different modes during a survey can affect how respondents' answer the same questionnaire.

Despite this possibility, the convenience and potential cost savings (especially relative to face-to-face/CAPI mode) have led Understanding Society to adopt a push-to-web mixed-mode design, starting at Wave 8, when 40% of participants were initially invited to complete the questionnaire online and a further 40% were initially approached for a face-to-face interview but then given the opportunity to complete online if they had not completed the face-to-face interview. The remaining 20% were only approached for a face-to-face interview. The implication of mode effects from Wave 8 onwards is that some of those people who chose web mode may have provided different answers to the same questions had they instead chosen CAPI. Given that 29% of Wave 8 individual interviews were carried out online, this means the introduction of mixed-modes could affect longitudinal analyses involving data from Wave 8 and earlier, predominantly CAPI, waves.

It is important to recognise that a substantively significant difference between the answers under web and under CAPI does not automatically imply that the web answer is ‘worse’. CAPI is only a benchmark for comparison with data from earlier CAPI-mode waves. D’Ardenne, Collins et al. (2017) discuss how mode effects depend on several features of how respondents answer survey questions (fear of disclosure, social desirability bias for sensitive questions and positivity bias, satisficing), and the presentation of the question and its possible answers, so which mode is ‘best’ will depend on the nature of each question.

Wave 8 involved an experiment in which a proportion of households in the first year were randomized to receive web first or CAPI first. The data from this experiment allow the estimation of the effect of web mode on key statistics in a way that takes into account that within the experimental sample the characteristics of those responding online and those responding by CAPI may differ. This training course shows how to use data from the experiment carried out at Wave 8 - Guidance on calculating the effect of mixed modes using Understanding Society.

We are currently investigating issues for users and will provide more detailed advice in due course. Unfortunately, it was not possible to devise a simple fix to adjust the results of every longitudinal analysis to equal what would have been obtained had those choosing web counterfactually chosen CAPI. Instead, we offer the following advice for those users who wish to investigate the impact of web mode on their analyses:

1. **Do not use the ‘indicator method’ for a regression/multivariable analysis**: The indicator method is simply to include a dummy variable that indicates whether the user chose web or CAPI as predictor variable in the regression analysis. However, despite its popularity, it was found that this approach is generally ineffective because it can often lead to badly biased results.

2. **A simple sensitivity analysis is to compare the estimates obtained using only the ring-fenced sample with those obtained using the remaining data**: The ring-fenced sample is a random sample of 20% of households for which the survey was administered CAPI-only, as in previous waves. The variable **h_ringfence** identifies members of this sample. To test whether the results of a regression analysis are different in the ring-fenced sample from

those in the mixed modes sample, the analyst can 1) include **h_ringfence** as a main effect in the model, and 2) include the interactions between **h_ringfence** and each predictor variable in the model. We recommend that the survey design and weights are accounted for when performing this analysis. If any of the interactions created in step 2 are statistically significant, this indicates the potential presence of mode effects. If the results are significant and you are unsure of how to proceed, it is recommended that you consult a statistician on your team to discuss.

Results of the experimental analysis for future waves will be added when ready.

Tips for analysts: COVID-19

With the arrival of COVID-19, all face-to-face interviews were suspended and we invited all our sample members to take part online or by telephone. Face-to-face interviews were used again from April 2022. We have brought together a document to help researchers explore Understanding Society changes to the main study due to the COVID-19 pandemic.

Support and resources

Understanding Society has a wealth of information for data users of all levels.

It is a highly comprehensive online source of information regarding its variables, methodology, survey design and implementation. It is also an up-to-date source of training courses, data releases and other relevant news regarding longitudinal research.

It's not a stupid question…

We are always pleased to hear from data users. You can contact us for help with using the data and for suggestions for improving the data.

Useful documentation links

As an introduction to the Understanding Society main study data and documentation we particularly recommend watching our 'getting started' videos and our pathway for new users page which explores the data and highlights the ways in which you can use the online resources to help you start using the data.

Videos

See our Exploring Understanding Society and Data Structure on our YouTube channel.

FAQs

Visit our frequently asked questions page for more information.

Index terms and Topic pages

Index Terms cover all the thematic areas in the Study and can be used to identify the variables most relevant to your research interests and to find other variables with related data throughout the dataset.

The topic pages provide an overview of each of the Study's thematic areas including information collected, the questions asked and links to related resources such as papers, webinars, blogs and news.

Syntax files for basic data management tasks and producing Derived Variables

A list of Stata commands for basic data management and analysis tasks can be found on the syntax web page.

Creating syntax with the Code creator

To help researchers get started with their research the Code creator extracts data from the Main Study and produces a simple flat data file. Select the variables needed from the Variable Search, 'save' the variables and 'build' the code. It provides you with ready-to-use Stata syntax to run on the downloaded data.

User Support

Help and Support for using Understanding Society can be found User support forum.

User support forum

After a short registration data users can read past issues, FAQs and report any issues or queries of their own. If you have a question about the data, post your question at the online data User forum. Users should read the 'How to raise an issue' guidance before posting a question. If you have a suggestion for improving the data, such as creating new derived variables, suggestions for data harmonisation, adding variables linked from external datasets, you can also post this via the User forum by selecting the category 'Suggestions for data improvements'.

The forum is monitored Monday-Friday and we aim to answer simple questions within 2 working days and more complex questions within 7.

Email

Users may also email User Support directly using our email address. Our preferred mode of communication is via the forum as other users may then also benefit from the information provided.

Online Helpdesk

If you'd like to speak to a member of the User Support team you can join an online helpdesk session. These are run via video conferencing software and are one-on-one sessions with a member of the User Support team. If you would like to access the online helpdesk please email us and we will respond with joining information and arrange a convenient time for the conversation.

Training courses

We offer both DIY training courses (via Moodle) and tutor-assisted workshops which give a general overview of the Study and demonstrate how to prepare the data for analysis using multiple statistical softwares.

Workshops are also available on specific aspects of the Study such as weights, biomarker and genetics data, the Innovation Panel, using the Study for transport analysis.

To learn more about these training workshops and how to register visit the training page.

Webinars and videos

Our webinars and training videos help inform data users about the Study and how to use the dataset. To explore our 'getting started' videos and past webinars visit our YouTube channel. For the latest list of webinars visit our website.

Email newsletter

Sign up to our newsletter for information about the Study.

Wave 14 content highlights

Wave 14 focussed on collecting an expansive set of data on families (including new entrants from the Wave 14 Boost sample) as well as new questions on long Covid. A summary of all changes to questions since Wave 13 is detailed in the Wave 14 Module summary table. The table lists the modules in the order they were asked within the Wave 14 questionnaire and includes when the modules were last asked.

The long-term content plan lists the Waves in which all the modules were last asked and when they will be asked again.

New modules for Wave 14 include:

- National Citizen Service (NCS)
- work illness
- generalised trust
- non-residential identifier
- non-resident children

For Wave 14 the Covid19 module focussed on questions about long covid to identify areas where the pandemic has had long term impacts on life. These are based on the monthly Covid-19 survey conducted as part of Understanding Society and allow further linkage across the two datasets.

The administrative data linkage consent modules asked in Wave 14 were last asked in Wave 11:

- HMRC
- National employment Savings Trust (NEST)

Most of the rotating modules for Wave 14 were last asked in Waves 11, 12 and 13:

- devolved election for - Scotland, Wales and Northern Ireland
- voluntary work
- domestic labour
- commuting behaviour
- work conditions
- transport behaviour
- family networks
- charitable giving
- social support
- physical work
- identity

The youth questionnaire for Wave 14 included rotating content on leisure - computer use, activities outside school. Family - household chores, supervision, friendship networks boy/girlfriend. Self-esteem. Risky behaviour - binge drinking, drugs, attitudes. Vandalism - fighting. Identity - ethnicity, religion. Future intentions -marriage, children, 10 years, future job.

Tips for analysts: COVID-19

In response to the pandemic the Wave 11 and 12 questionnaires were adapted to capture changes during this time. Updates went into the field on 28 July 2020. Wave 13 included some questions from the Covid-19 Survey. Wave 14 included questions about long covid (based on the Covid-19 Survey) to identify areas where the pandemic has had long term impacts on life. Researchers need to take into account how the pandemic impacted the main study and consider the effects of mode transition, changes to the questionnaire and analysis of changes during the pandemic compared to the pre- and post pandemic. We have brought together a document to help researchers explore Understanding Society changes to the main study due to the COVID-19 pandemic.

COVID-19 dataset

The COVID-19 survey started in April 2020 in response to the COVID-19 pandemic and interviewed participants from the main Understanding Society sample via a web-survey. This started as a monthly survey and shifted to bimonthly after July 2020 and continued until September 2021. In addition to questions directly related to Covid-19 (symptoms, testing, vaccination), the survey includes questions on different aspects of people’s lives that could have been impacted by the pandemic. The released data also includes the data on serology antibody testing conducted in March 2021 and 2019 pre-pandemic data from the main survey interviews.

Revisions to previous releases

Each time we release a new wave or new edition of data, we include all previous waves.

Users can find information in the UKHLS 2024 Revisions document, supplied with the study documentation downloaded from the UK Data Service.

To enhance the data we would like to hear from researchers about any errors, inconsistencies, or other problems identified when using the data. Please contact our Data User Support service with any issues relating to data or data analysis.

Revisions to previous Waves from Wave 5 onwards are included below:

UKHLS 2023 Revisions Waves 1-12 document

UKHLS 2022 Revisions Waves 1-11 document

UKHLS 2021 Revisions Waves 1-10 document November

UKHLS 2020 Revisions Waves 1-9 document

UKHLS 2019 Revisions Waves 1-8 document

UKHLS 2018 Revisions Waves 1-7 document November

UKHLS 2018 Revisions Waves 1-7 document July

UKHLS 2017 Revisions Waves 1-6 document

UKHLS 2016 Revisions Waves 1-5 document