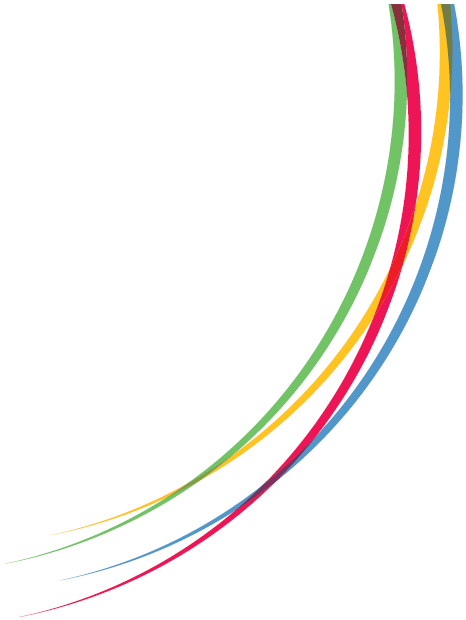


Understanding Society

The UK Household Longitudinal Study

Waves 1-8

User Guide

Edited by

Gundi Knies

Institute for Social and Economic Research

University of Essex

Colchester

Essex

November 2018

CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

LIST OF ABBREVIATIONS

BHPS	British Household Panel Survey
HBAI	Households Below Average Income Statistics
CAPI	Computer Assisted Personal Interview
CASCOT	Computer Assisted Structured Coding Tool
CASI	Computer Assisted Self-Interview
CNEF	Cross-National Equivalent File
DWP	Department for Work and Pensions
ECHP	European Community Household Panel
EMB	Ethnic Minority Boost
ESRC	Economic and Social Research Council
EUL	End-user Licence
GOR	General Office Region
GPC	General Population Comparison
GPS	General Population Sample
HMRC	Her Majesty's Revenue and Customs
ICE	Imputation by chained equations
IP	Innovation Panel (UKHLS component for methodological research)
IEMB	Immigrant and Ethnic Minority Boost
ISER	Institute for Social and Economic Research
LDA	Low density ethnic minority area
LSOA	Lower Layer Super Output Area
MSOA	Middle Layer Super Output Area
NatCen	National Centre for Social Research
NIHPS	Northern Ireland Household Panel Survey
NISRA	Northern Ireland Statistics and Research Agency
NI	Northern Ireland
ONS	Office for National Statistics
OSM	Original Sample Member
PMM	Predictive mean matching
PSM	Permanent Sample Member
SIC	ONS Standard Industry Code
SL	Special Licence
SOA	Super Output Area
SOC	ONS Standard Occupational Classification
TSM	Temporary Sample Member
UKDS	UK Data Service (previously: UK Data Archive/ UKDA)
UKHLS	UK Household Longitudinal Study (official acronym for *Understanding Society*)

1. INTRODUCTION

1.1. WHAT IS UNDERSTANDING SOCIETY?

Understanding Society, the UK Household Longitudinal Study (UKHLS), is a longitudinal survey of the members of approximately 40,000 households (at Wave 1) in the United Kingdom, i.e., the geographical area of the countries England, Scotland, Wales and Northern Ireland (NI). Households recruited at the first round of data collection are visited each year to collect information on changes to their household and individual circumstances. Interviews are typically carried out face-to-face in respondents' homes by trained interviewers. From Wave 3 onward, a small number of respondents are interviewed over the phone and from Wave 7 onward some proportion of the sample provides their information in a web interview. Data collection for each wave takes place over a 24-month period. Note that the periods of waves overlap, and that individual respondents are interviewed around the same time each year.

Understanding Society is funded by the Economic and Social Research Council (ESRC) and with funding from multiple government departments (the Department for Work and Pensions (DWP), the Department for Education, the Department for Transport, the Department for Culture, Media and Sport, the Department for Communities and Local Government, the Department of Health, the Scottish Government, the Welsh Assembly Government, the Northern Ireland Executive, the Department for Environment, Food and Rural Affairs, and the Food Standards Agency). The scientific leadership team is from the Institute for Social and Economic Research (ISER) of the University of Essex, the University of Warwick, and the London School of Economics. Professor Nick Buck was the principal investigator until June 2015. Professor Michaela Benzeval has been the principal investigator since July 2015. Fieldwork was conducted by the National Centre for Social Research (NatCen) with collaboration with the Central Survey Unit of the Northern Ireland Statistics and Research Agency (NISRA) in Northern Ireland (Waves 1 to 5) and by TNS BMRB (now known as Kantar Public), with collaboration with Millward Brown Ulster in Northern Ireland (Waves 6 to 9).

The overall purpose of *Understanding Society* is to provide high quality longitudinal data about subjects such as health, work, education, income, family, and social life to help understand the long term effects of social and economic change, as well as policy interventions designed to impact upon the general well-being of the UK population. To this end, the Study collects both objective and subjective indicators and offers opportunities for research within and across multiple disciplines such as sociology and economics, geography, psychology and health sciences. The Study also provides a platform for additional data collections.

1.2. HOW TO NAVIGATE THIS USER GUIDE

This release has data for the *Understanding Society* main study which collects information from the UK General Population Sample (GPS) and the Ethnic Minority Boost Sample (EMBS). From Wave 2 onward the main study also includes information collected from continuing participants of the British Household Panel Survey (BHPS), a household panel survey of around 8,000 households in the UK,

which has completed 18 annual waves of data collection and has been run by ISER since it began in 1991. To learn more about the BHPS and other components of *Understanding Society*, see Section 5.3, below. From Wave 6 onward the main study also includes an Immigrant and Ethnic Minority Boost Sample (IEMBS). From the Wave 7 data release (November 2017) onward, releases also include *Understanding Society*-harmonised BHPS data (henceforward: harmonised BHPS). Users interested in this element of the Study should consult the designated *Understanding Society* harmonised BHPS User Guide (Fumagalli, Knies et al. 2017) in addition to reading this guide, which focuses on using the main Understanding Society data.

The User Guide is structured as follows. We first present the general aspects of the Study design (Section 2), which covers sample design (Section 2.2), data collection (Section 2.3), data processing (Section 2.3.4.8), and questionnaire content (Section 2.5). Section 3 provides a description of *Understanding Society* data files and variables, covering derived variables (Section 3.2.6), weighting adjustments (Section 3.3), derived income variables (Section 3.4) and example code for matching information contained in different files (Section 4.7). Information on how to access the data is provided in Section 5. Within it, Section 5.3 provides additional information about further studies in the *Understanding Society* family such as the stand-alone BHPS, the *Understanding Society* Nurse Health Assessment, the Innovation Panel and the Cross-National Equivalent File.

As an introduction to the *Understanding Society* main study data and documentation we particularly recommend the following reading:

- The summary of the general questionnaire content (Section 2.5.3), and notes on naming conventions (Section 3.2.2),
- The sections on sample design (Section 2.2), weighting adjustments (Section 3.3) and data collection and response outcomes (Section 2.3).
- Variable level descriptions of the data can be found on the Study website (https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation). The online documentation has extensive links between questions and detailed views of variables and data files. There is also a search facility for searching questions, variables, modules, and data files.
- The example Stata code for matching variables from different data files (Section 4.7).

In assembling the documentation, we have drawn upon the documentation for the BHPS, see Taylor (2010) and http://www.iser.essex.ac.*uk*/bhps.

2. UNDERSTANDING SOCIETY STUDY DESIGN

2.1. OVERVIEW

Understanding Society is a panel survey of households with yearly interviews. Data collection for a single wave is scheduled across 24 months. The Study began with a representative probability sample of households. There is an extended discussion of sample design in Section 2.2, and in Lynn (2009). Adult household members (age 16 or older) are interviewed and the same individuals are re-interviewed in successive years to see how things have changed. Household members aged 10-15 years are asked to complete a short self-completion youth questionnaire. Children become

eligible for a full interview once they reach the age of 16. We refer to them as “Rising 16s”.

The overall Study has multiple sample components. In the main survey there is

- the General Population Sample (GPS), with its subset the General Population Comparison (GPC) sample,
- the Ethnic Minority Boost Sample (EMBS),
- the BHPS sample from Wave 2 onward, and
- the Immigrant and Ethnic Minority Boost Sample (IEMBS) from Wave 6 onward.

All samples are administered the same survey instruments and asked the same questions with some exceptions: Some sample members (see Section 4.5 for further details about the constituents of this sample) are asked an additional “Extra 5 minutes” worth of questions that are particularly relevant for ethnic minority and immigrant communities (e.g., ethnic identity and remittances). Additionally, at Wave 6, the instrument for the then new IEMB sample was similar but not identical to the questionnaire administered to the other samples (see Section 2.5.3.6 for further details).

The instruments for the first three components are the same except the EMBS, IEMBS and the GPC sample have an “Extra 5 minutes” of questions specifically relevant to ethnic minority communities (e.g., ethnic identity and remittances). At Wave 6, the instruments for the new IEMBS were very similar but the focus was on collecting information relevant to new entrants, immigrants and ethnic minorities.

Data from 18 waves of the BHPS have been included for the first time with the *Understanding Society* Wave 1-7 data release, in November 2017. The data draw on data published in 2009 (and available as UK Data Service SN5151) but include *Understanding Society*-harmonised variable names, cross-wave identifiers that work across the two studies and much more. Basic documentation is provided in this guide but users should also consult Fumagalli, Knies et al. (2017).

In Waves 2 and 3, *Understanding Society* augmented survey questions with direct health assessments and the collection of blood samples. The Health Assessment data can be accessed through the UK Data Service (UKDS), SN7251. Documentation is provided separately, see McFall, Petersen et al. (2014).

In addition, there is a separate survey, the Innovation Panel (IP), which is fielded in the year before the main survey. It tests varying measurement issues, and its instruments are somewhat different from the main survey. The IP can be accessed through the UKDS, SN 6849. Documentation is provided separately, see Al Baghal, Jaeckle et al. (2015).

2.2. SAMPLE DESIGN

The *Understanding Society* main survey sample consists of a new large General Population Sample (GPS) plus three other components: the Ethnic Minority Boost Sample (EMBS), the former BHPS sample, and the Immigrant and Ethnic Minority Boost Sample (IEMBS). The design of the first three components is described in more detail in an *Understanding Society* working paper, see Lynn (2009). The design of the IEMBS is described in Lynn, Nandi et al. (2016). The GPS is based on two

separate samples of residential addresses in England, Scotland and Wales and in Northern Ireland. The England, Scotland and Wales sample is a proportionately stratified (equal probability), clustered sample of addresses selected from the Postcode Address File. Northern Ireland has an unclustered systematic random sample of addresses selected from the Land and Property Services Agency list of domestic addresses.

2.2.1. GENERAL POPULATION SAMPLE

The sample for England, Scotland and Wales was selected in two stages. The first stage was to select a sample of postcode sectors as the primary sampling units (PSU's). The second stage was to select addresses within each sampled sector. Prior to selection, any postcode sector with fewer than 500 residential addresses was grouped with an adjacent sector and thereafter treated as a single sector. The list of all sectors was then sorted into twelve geographical strata, consisting of ten regions in England plus Scotland and Wales as separate strata. Within each of the twelve strata, sectors were sorted into three sub-strata based upon the proportion of household reference persons classified as non-manual workers, from 2001 Census data. Within each of the 36 sub-strata, sectors were then sorted into three further sub-divisions based on population density (households per hectare) and within each of the 108 resultant sub-divisions, sectors were listed in order of ethnic minority density. From the sorted list, a systematic random sample of 2,640 sectors was selected, with probability proportional to the number of residential addresses in the sector. These sectors were then allocated systematically to 24 monthly samples, with 110 sectors in each monthly sample. Within each postal sector, 18 addresses were selected using systematic random sampling. The England, Scotland and Wales sample in this data release is based upon an initial sample of 47,520 addresses.

In Northern Ireland, 2,395 addresses were selected in a single stage from the list of domestic addresses. In combination, this data release is therefore based upon a total of 49,915 addresses.

At each address, the final stage of sampling was carried out by field interviewers. This consisted of identifying persons to be defined as sample members. All persons resident at each sample address at the time the interviewer made contact were deemed to be a sample member, with the exception of the small proportion of addresses that contained more than three dwellings or households. In those cases, three dwellings or households were sub-sampled at random.

2.2.2. GENERAL POPULATION COMPARISON SAMPLE

The General Population Comparison sample (GPC) has one sampled address for 40% of the selected postal sectors in General Population Sample (GPS) component for Great Britain. In other words, of the 2,640 general population sectors, 60% of them (1,584) contain 18 GPS addresses and the other 40% contain 17 GPS addresses and one GPC address. The persons in these households will be designated as members of the GPC sample, regardless of ethnic group membership. Members of the GPC sample are a random subsample of the GPS component and they should be included in analyses of the GPS component.

2.2.3. ETHNIC MINORITY BOOST SAMPLE

The EMBS was designed to provide at least 1,000 adults from each of five groups: Indian, Pakistani, Bangladeshi, Caribbean, and African.

The initial step was identifying postal sectors with relatively high proportions of relevant ethnic minority groups, based upon 2001 Census data and more recent Annual Population Survey data. The set of 3,145 sectors constituted approximately 35% of the sectors in Great Britain and covered between 82% and 93% of the population of the five ethnic minority groups.

The 3,145 sectors were sorted into four strata based on the expected number of ethnic minority households that would be identified by the sampling and screening procedures (see Berthoud et al., 2009 for details). All sectors were included for the stratum where a yield of three or more households was expected. In the other three strata, sectors were sub-sampled at rates of one in four, one in eight, or one in 16, respectively. This was done to constrain the number of sectors that might have just one or two eligible sample households (or even none). The total number of postal sectors selected for inclusion in the EMBS was 771. Of these six were in Scotland, seven were in Wales, and the remaining 758 were in England, with a concentration in London (412 sectors).

The number of addresses selected per postal sector ranged from 15 to 103. Sampling fractions varied across the sectors in a way designed to deliver target numbers of respondents in each target ethnic minority group with adequate statistical efficiency (see Berthoud, Fumagalli et al. (2009) for more details). In sectors selected for both the GPS component and the EMBS, a single systematic sample of the required total number of addresses was selected and allocated in a systematic way to the two sample components, thus ensuring that both sample components are spread throughout the whole sector.

The final stage of sampling was done by the interviewers. The steps are described in the Project Instructions for Interviewers. At addresses containing more than three dwellings or households, the procedures to sub-select dwellings or households were as described above for the GPS component. Within each household, rather than all resident persons becoming sample members, there were three additional steps:

- A “screen” was carried out to identify whether there were any persons from target ethnic groups in the household.
- A random mechanism was applied to certain target groups identified by the screen in order to select only a desired proportion into the sample (non-mixed Indian, Pakistani, non-mixed Caribbean, African, Far Eastern, Middle Eastern). For other target groups, all resident persons were included in the sample (mixed Indian, Bangladeshi, mixed Caribbean, Sri Lankan, Chinese, Turkish).
- In households included in the sample in the previous two steps, all members of target ethnic groups were deemed to be members of the EMBS (including children). All persons of other ethnic groups are not EMBS members. They will be interviewed as temporary sample members for so long as they remain co-resident with at least one EMBS member.

The overall sampling fractions combine a) the probability of sampling the sector, b) the fraction of addresses selected within the sector, and c) the probability of a

household being retained following the application of the random selection mechanism described above.

2.2.4. FORMER BHPS SAMPLE

The sample issued at Wave 2 consisted of all members from the BHPS sample who were still active at Wave 18 of the BHPS and who had not refused consent to be issued as part of the *Understanding Society* sample. It should be noted that the BHPS sample contains different components, including the original sample (first selected in1991), boost samples in Scotland and Wales (first selected in 1999), and a Northern Ireland sample (selected in 2001). For further details of the BHPS sample, see Section IV of the BHPS User Guide (Taylor 2010).

2.2.5. IMMIGRANT AND ETHNIC MINORITY BOOST SAMPLE

This sample was introduced at Wave 6. It includes people who were born outside the United Kingdom ("immigrants") and members of five ethnic minority groups: Indian, Pakistani, Bangladeshi, Caribbean, and African. Some people, of course, fall into both categories. This sample therefore provides coverage for the first time of people who have entered the UK since Wave 1 of the Study ("new immigrants"), while also boosting the numbers of immigrants who arrived earlier and of ethnic minorities who either arrived earlier or were born in the UK. The IEMBS was designed to provide around 2,000 adult immigrant respondents and around 2,500 from the target ethnic minority groups.

The sample was identified through in-person doorstep screening of a set of addresses that were sampled from the Postcode Address File following a stratified multi-stage design in which the strata were defined by small area level indicators from the 2011 population census of the distribution of ethnic groups and immigrants. Five strata were created. Sampling was restricted to four strata, the fifth consisting of the sectors with the very lowest proportions of immigrants and ethnic minorities. Sampling fractions varied between the four strata, with the highest sampling fraction applied to a stratum with the highest proportions of Africans. In each sampled stratum, a number of postcode sectors were selected with probability proportional to the predicted number of eligible households. In each sampled sector, a number of addresses were selected such that the predicted number of eligible households in the sample did not vary between sectors within a stratum (so the number of selected addresses was larger in sectors with a lower predicted proportion of eligible households). A screened household was eligible for interview if it contained at least one person who was born outside the UK and/or a member of a relevant ethnic minority group, even if that person was a child.

The "boost" samples do not therefore provide complete population coverage of the relevant subgroups but are instead designed to be used in combination with the other samples, as described above. The sample of "new immigrants" is estimated to provide around 74% population coverage.

2.2.6. SAMPLE STATUS AND FOLLOWING RULES

There are three possible sample statuses: Original Sample Members (OSMs), Temporary Sample Members (TSMs), and Permanent Sample members (PSMs).

2.2.6.1. Original Sample Members (OSMs)

All members of *Understanding Society* GPS households enumerated at Wave 1 - including absent household members and those living in institutions who would otherwise be resident - are Original Sample Members (OSMs). All ethnic minority members of an enumerated household eligible for inclusion in the EMBS are OSMs. In the IEMBS, each household member who met the eligibility criteria at Wave 6 was deemed an OSM.

In all of these samples, any child born to an OSM mother after Wave 1 and observed to be co-resident with the mother at the survey wave following the child's birth is an OSM. In the former BHPS sample, OSMs are those who were enumerated at the first wave of the sample from which they come (Wave 1 for the original sample, Wave 9 for the Scotland and Wales boost samples, Wave 11 for Northern Ireland) or who were subsequently born to an OSM mother or father (or both). Following the incorporation into *Understanding Society* from Wave 2 onward, in the former BHPS sample, as for all other *Understanding Society* samples, only children born to an OSM mother will themselves become an OSM. OSMs, of all ages, are followed for interview and remain eligible as long as they are resident within the UK. They remain potentially eligible sample members for the life of Study.

The case may arise where the only OSM in the household is a child. Other household members are then TSMs so long as they are co-resident with the child, and therefore eligible for interview, even if the child is not yet old enough to be eligible for interview. If the OSM child moves house, they are followed to their new address and those living with the OSM child are eligible for interview. If the OSM child moves into an institution, where normally just the OSM/PSM would be interviewed and not co-residents, a split-off household is created containing only the OSM child and the household enumeration grid completed. The child OSM is an eligible sample member, even if they are not eligible for interview because of their age.

2.2.6.2. Temporary Sample Members (TSMs)

Any members of an enumerated household eligible for inclusion in the EMBS at Wave 1 who are not from a qualifying ethnic minority are Temporary Sample Members (TSMs) at Wave 1. This was the only category of TSM at Wave 1. Likewise, any members of an enumerated household eligible for inclusion in the IEMBS at Wave 6 who do not have a qualifying ethnic minority or immigration background (non-ethnic minorities who were born in the UK) were deemed to be TSMs at Wave 6.

In all parts of the sample, any new person found to be co-resident in an OSM or PSM household after Wave 1 is a TSM. This would include any child born to an OSM father after Wave 1 but not an OSM mother and observed to be co-resident with the father (or any other OSM) at the survey wave following the child's birth. TSMs remain eligible for interview as long as co-resident in an OSM/PSM household. TSMs who are not co-resident in an OSM/PSM household are not followed and become ineligible for interview. TSMs are identified as re-joiners if they are subsequently found in an OSM/PSM household and then become eligible for interview.

2.2.6.3. Permanent Sample Members (PSMs)

PSMs are TSMs who are followed for interview after they no longer live with an OSM. This is done for substantive research reasons because of the additional contextual information they may provide for the analysis of OSMs. At present, there is only one category of PSM, but others may be defined in the future. Any TSM father of an OSM child born after Wave 1 and observed to be co-resident with the child at the survey wave following the child's birth is a PSM. PSMs remain potentially eligible for interview for the life of survey.

2.3. DATA COLLECTION AND RESPONSE OUTCOMES

2.3.1. OVERVIEW

Understanding Society is issued to field as 24 monthly samples. There is some variation in this pattern. The Northern Ireland and the former BHPS sample components are issued over the first 12 months of the wave, and the IEMB is issued over the second 12 months of the wave. Table 1 shows the timing of the sample issue for the data included in this release.

Most of the data collection is conducted face-to-face via computer aided personal interview (CAPI). There are also self-completion instruments for youth and adults. The youth instruments are administered on paper. The adult self-completion questionnaire shifted from paper to computer administered self- interview (CASI) in Wave 3. From Wave 3 onward, there was also a telephone mop-up at the end of the fieldwork period for each sample month. At Wave 7, online questionnaires were introduced for the first time, but only for households in which no member had participated at Wave 6. From Wave 8 onward, online interviewing is a core data collection mode.

2.3.1.1. Mixed mode data collection (from Wave 8 onward)

From Wave 8 adults in a proportion of those households which *had* responded at Wave 7 were also invited to take part online. Using data from the Innovation Panel, a model was developed which predicted the propensity of each household completing all the eligible interviews online, whilst minimising the risk of the household members refusing an in-person interview if invited to take part online first. This model was then used to allocate households who were invited to take part online-first at Wave 8. Thus, in the main-sample of *Understanding Society*, unlike the Innovation Panel, the allocation to initial mode is not random. It is those households that we think are most likely to take part online who are invited to take part online. However, a random 20% of all *Understanding Society* households have been designated as a ring-fenced CAPI-only sample. The procedures for allocating some households to be invited to take part online were applied only to the remaining 80%.

Table 1: Timing of data collection start

Year		2008				2009				2010				2011				2012				2013				2014				2015				2016				2017				2018			
Quarter		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BHPS Wave 18																																													
UKHLS	W1																																												
	W2																																												
	W3																																												
	W4																																												
	W5																																												
	W6																																												
	W7																																												
	W8																																												

Notes: Northern Ireland (from Wave 1 onward) and BHPS (from Wave 2 onward) samples interviewed in year 1 of each wave only. IEMB (from Wave 6 onward) sample interviewed in year 2 of each wave only. Note that fieldwork typically continues for an additional 5 months to interview eligible sample members who did not provide interviews when first issued to field. We only have access to complete wave information in the third quarter following the official conclusion of fieldwork for the wave (but some data processing can start early, for some proportion of the sample).

development
data collection
data processing & documentation
data release

Those sample members who did not complete online in the first few weeks were then issued to interviewers, who would try to contact and interview them in person, although the sample members could still access the online interview. This group is referred to as the "web-first" group, because they are first invited to take part online and then followed up in other modes. At this point, those households that had not been issued to web were also issued to interviewers. This group of households are the "CAPI-first" households, because they are first issued to interviewers and non-respondents are followed up in other modes. During the re-interview phase, non-responders in the CAPI-first households were invited to take part online, although interviewers still attempted to get an in-person interview. Both web-first and CAPI-first non-responders were also eligible for the telephone mop-up towards the end of the fieldwork period.

Overall, at Wave 8, 20% of households were issued CAPI-only, 40% CAPI-first and 40% web-first. Amongst the 80% that were not ring-fenced to be CAPI-only, those that had not responded at wave 7 were issued web-first and those households with the lowest propensity to respond by web were issued CAPI-first. During the first year of fieldwork (first four sample quarters) the remaining sample was randomly allocated between web-first and CAPI-first so as to produce the desired overall distribution by modes. In the second year, rather than a random allocation the lowest propensity households were all issued CAPI-first and the remainder all web-first.

During the first year of Wave 8 data collection, we also implemented an adaptive 'push-to-web' design to experiment with ways to increase the proportion of households in which all eligible adults took part online before remaining households were issued to an interviewer (for more details, see Carpenter and Burton 2018). The design, which was carried over into the second year, consisted of an invitation letter to each adult in a web-first household, which included details of how to access their unique survey online. The invitation letter also included the incentive. For those adults for whom we had an email address, an email was sent with a direct link to the survey. After one week a reminder letter and email was sent, with another letter and email sent one week after that. For those who had not yet responded online, another letter and email was sent after three weeks of the original, with a reminder email after four weeks. The online only fieldwork finished after five weeks, and those who had completed their survey online by that point were sent a bonus £10 gift voucher. For the CAPI-first sample, the process was similar to previous waves, with an advance letter (containing the incentive) being sent before interviewers called. Adults that were re-issued received the re-issue letter, which at Wave 8 included a unique link to complete online.

2.3.2. DATA COLLECTION

2.3.2.1. The players, who does what

ISER and the fieldwork agencies work closely together on all aspects of data collection, implementing an agreed set of survey procedures designed to ensure adequate response and effective data quality. For Waves 1 to 5, the agencies responsible for data collection were NatCen Social Research (Great Britain), and the Central Survey Unit of NISRA (Northern Ireland). For Waves 6 to 8, the fieldwork in Britain is conducted by TNS BMRB (now Kantar Public), with the Northern Ireland fieldwork being completed by Millward Brown Ulster.

ISER has the primary responsibility for design work. The fieldwork agencies manage fieldwork, editing, coding and data-entry. They also advise on the design of all research instruments. ISER plays a major role in quality control through specification of fieldwork practices, survey materials, editing and coding requirements, and monitoring and analysing weekly fieldwork progress reports. This working relationship is reinforced by an agreed set of survey-specific procedures to ensure adequate response and effective data quality. Full details of these, and other technical aspects of the data collection and fieldwork, coding, and data processing are found in the *Technical Reports,* published each wave on the *Understanding Society* website (see https://www.understandingsociety.ac.uk/documentation/mainstage/technical-reports). The *Understanding Society* Quality Profile reports additional aspects of survey and data quality, including of the data released to researchers (Lynn and Knies 2015).

2.3.2.2. Getting Ready for Fieldwork

Prior to the first wave of the *Understanding Society*, there were two small pilot studies and a dress rehearsal. A cognitive pilot of 70 individuals was conducted March – April 2008 to test screening and other questions relevant to the ethnicity strand. A translation pilot was conducted in June 2008: 50 interviews were carried out using Bengali and Punjabi translations of the questionnaire to see if there were problems with the operation of the translation program or problems with interviewing with the translated instruments. A run-through of all data collection instruments and procedures in 100 households, called a dress rehearsal, took place August-September 2008.

A pilot for Wave 2 tested all instruments and data collection procedures. For this wave, the data collection also focused on assessing any problems with integrating members of the former BHPS sample component, which includes a small segment conducted by telephone interviews. In all 237 households were issued. Of these, 91 were households interviewed in the Wave 1 pilot. The BHPS sample component was represented by households that were part of the BHPS between 1997 and 2001, the European Community Household Panel (ECHP). Households for which we had a telephone number were issued to telephone interview to test the telephone interview instruments and procedures. The Wave 2 pilot took place September-October 2009.

Subsequent pilots and dress rehearsals also returned to these samples and took place in September to November, leaving up to two months to address any issues before the start of the main wave.

2.3.2.3. Interviewers

We have tried to use interviewers of above average levels of experience and ability because of the demanding nature of *Understanding Society***.** The majority of interviewers in Northern Ireland had worked on the BHPS Northern Ireland component (the Northern Ireland Household Panel Survey), and were familiar with the design and operation of *Understanding Society*.

In addition to general interviewer training, interviewers working on the Study attended a one day survey-specific briefing. Generally around 12 to 20 interviewers attended each briefing, along with two or three briefing managers or area managers. The briefings were led by at least one researcher from NatCen with the majority also attended by ISER staff. The briefings in Wave 1 took place across the UK; Belfast,

Birmingham, Brentwood, Bristol, Derby, Edinburgh, Glasgow, Leeds, London and Manchester. Similar topics and locations were used for the Wave 2 briefings. At Wave 3, the Edinburgh briefing was dropped and two briefings were held in Glasgow. Additional briefings were added in Bury St. Edmonds, Liverpool and Gateshead.

The morning sessions were devoted to fieldwork procedures, for example the administrative forms to record contact information, and how to deal with the complexities of multiple dwelling units and multiple households. The afternoon was spent discussing the survey content and reviewing and working with the Blaise computer aided personal interview (CAPI) instrument. At Wave 3, there were two types of briefing; for interviewers experienced with the Study or for interviewers with experience who were new to the Study. The latter briefing went into more detail about the background of *Understanding Society*, early findings, the more technical details of the sample, and the task of enumerating the household.

At Wave 4, the style of briefing changed in Great Britain. Interviewers had worked on the Study for three waves and were familiar with the mechanics of how to conduct the survey. For those interviewers returning for Wave 4, the focus of the briefing switched from the survey procedures to motivating the interviewers and giving them information to enable them to motivate the sample members when making contact. Interviewers who were new to the survey still attended a standard briefing, as did interviewers in Northern Ireland. These standard briefings were held in Belfast (3) and London (1).

Experienced interviewers attended ‘conference-style’ briefings. These briefings were much larger than standard briefings, with 150-250 interviewers attending each event. There were three such events held prior to the start of Wave 4; in Birmingham, Liverpool and London. During the breaks in day, there were stalls and displays of media coverage, research findings, and information about the Study, a Twitter stand and an area where interviewers could write questions on post-it notes for discussion later in the day. The content of the briefing consisted of ‘plenary’ sessions where an overview of progress on the Study “so far” was presented, along with researchers from ISER or the LSE talking about how they used *Understanding Society* in their research, videos of the Chief Executive of NatCen (Penny Young) and the then ISER Director (Professor Heather Laurie MBE) were shown, and medals awarded to interviewers who had achieved 100% response rate in any of their allocations at Wave 3. Once during the morning, and once in the afternoon, there were a number of ‘break-out’ sessions with small groups of interviewers to share best practice and experience of (i) contact and co-operation and (ii) how to deal with household splits and allocating outcome codes. The discussions of these break-out sessions were then discussed at the plenary sessions.

The Wave 5 briefings were also conducted as ‘conference-style’ briefings, with four such briefings; two in London, one each in Glasgow and Liverpool. In addition, there were three standard briefings in Northern Ireland (all in Belfast). The structure was similar to that in Wave 4, with an introduction from the Chief Executive of NatCen Social Research followed by presentations covering what was new for Wave 5, results of qualitative research carried with on *Understanding Society* sample members, former-sample members and some interviewers. Results from a recent Innovation Panel were shared, with information on how this has impacted on the

design of Wave 5. Also, an example of quantitative research using *Understanding Society*, looking at religion, was presented by a NatCen researcher.

Interviewers were assigned to specific areas. For Wave 1, 911 interviewers were employed to cover 3,517 areas in the sample. The number of interviewers briefed in Wave 2 was 819 and 746 at Wave 3. At Wave 4, 692 interviewers worked on the Study. At Wave 5, 570 interviewers worked on the Study.

At Wave 6, with the change in the responsible fieldwork agency, there was a need to return to full briefings with smaller groups. Briefings started in London on December 4th, 2013. During December there were in total 10 briefings, held in London, Warwick, Bristol, Edinburgh, Exeter, Newcastle, Cambridge, Manchester and Belfast. Initially, interviewers were briefed who would be working in the first couple of sample months. Additional briefing sessions were held throughout 2014, to bring interviewers on to the project who were starting work later on in the year. At Wave 6 there were 564 interviewers working on the Study.

At Wave 7, there were a mixture of full day briefings (for interviewers new to the study) and half day 'refresher' briefings for those who were working on Wave 6. The first Wave 7 briefing was in London on the 8th December. In total, there were eleven briefings in December, held in London, Manchester, Warwick, Bristol, Cambridge, Newcastle, and Belfast. Additional briefings were held throughout the year, to coincide with the start of each quarter of fieldwork. Across Wave 7, there were 399 interviewers working on the study.

At Wave 8, the format of the briefings was similar to Wave 7, with a mixture of full day and half-day 'refresher' briefings. There were 30 briefings in total, held in Belfast, Bristol, Glasgow, Leeds, London, Manchester, Newcastle, and Nottingham. Altogether, 345 interviewers worked on the study.

2.3.2.4. Fieldwork

When beginning fieldwork for Wave 1, we did not know who was in the sample. Interviewers mailed an introductory card from ISER to all sampled addresses (addressed to "The Occupier"), together with a small leaflet outlining the purpose of the survey. Then the interviewer called within a week of the mailing. At the end of the first interview, all participating households received a more detailed brochure, giving further information about the survey and thanking respondents for participating.

A minimum of six calls is made at each sampled address before it is considered a non-contact. Interviewers are encouraged to make further calls, if possible. If there is a potential for success, a special conversion letter is sent to households which had refused to participate or had not been contacted. Post interview quality control is carried out with a telephone recall on 10% of all completed interviews.

Interviewers upload their work daily, including information about all the calls they have made, whether or not there was any response. This information is collated by NatCen to construct a weekly field progress monitor report for ISER.

During the second year of Wave 3 (2012) a telephone "mop-up" was introduced. This was started in April, but also covered the sample from January-March. The aim of this was to contact adults who could not be contacted by face-to-face interviewers during the main fieldwork period. Adults in households that were non-responding in the main fieldwork period, except those who had adamantly refused or were deemed

to be mentally or physically incapable of participating, were contacted by the NatCen Multi-Mode Unit based in Brentwood. The trained and briefed telephone interviewers at the Multi-Mode Unit introduce themselves, remind the sample members of the survey, and ask whether they would be able to do the interview by telephone. The purpose of the mop-up" was to increase participation among those who were hard to contact in person.

Analysis by NatCen indicates that the telephone mop-up increased the overall household response rate for that period by about three percentage points for the EMBS and by just less than two percentage points for the GPS. This mop-up was not conducted with the BHPS sample in Wave 3 since they are interviewed in the first year of each wave.

Towards the end of Wave 3, September 2012, a trial was conducted in two field areas in which an additional incentive was used at the re-issue stage. This was then rolled-out across the sample from October, and so covers the last quarter of Wave 3. In the implementation, non-responding households were reviewed by the NatCen Operations Department in Brentwood for re-issue and possible re-allocation to a different interviewer. Households which had refused to participate in the initial fieldwork period, but where the assessment was that this was a "soft" refusal, were sent a re-issue letter which mentioned an additional incentive if they participated during the re-issue fieldwork period. Other non-responding households were sent a normal re-issue letter, but the interviewers had discretion to offer the additional incentive on the door-step if they felt that this would convert a non-responding household to a participating household.

In addition, during the latter quarter of Wave 3 fieldwork, more effort was made to increase interviewer continuity for households across waves, rather than prioritising interviewer efficiency. It is estimated that these two procedures, which were launched almost simultaneously, increased household response rates by around four percentage points for the EMBS and by around two and a half percentage points in the GPS in Quarter 8. The procedures adopted in Wave 3 to maintain household response were continued in the fieldwork at Waves 4 and 5.

The contract for fieldwork for Waves 6 to 8 was awarded to TNS BMRB (now Kantar Public), with Millward Brown Ulster conducting the fieldwork in Northern Ireland. The fieldwork procedures for the Study remained fairly consistent. There were some minor changes, however. The incentive increased to £20 for adults in households that had not participated at Wave 5. Where the household had refused, the incentive was unconditional and included with the advance letter. Where the household had not been contacted, or were non-responding for other reasons, the incentive was conditional on participation at Wave 6. This design replaced the use of additional incentives at the re-issue stage. Telephone interviews were no longer done at a central telephone unit, but were done by the face-to-face interviewers, but using a CATI-version of the script and calling from their own homes. From an administration perspective, the sample was managed electronically. Rather than using paper "Address Record Forms", the interviewers used an Electronic Contact Sheet, which was part of the sample management process on their lap-tops.

2.3.3. PANEL MEMBERSHIP AND PANEL MAINTENANCE

The rules for following individual respondents over time are based upon the composition of the household. Individuals found at selected households in the first wave were designated as Original Sample Members (OSM). We attempt to maintain OSM respondents as part of the sample as long as they live in the UK. In addition, births to an OSM mother are classified as OSM. Individuals joining the household of an OSM after enumeration of the household at Wave 1 are Temporary Sample Members (TSM). One deviation from this is for individuals who were not an ethnic minority within the households selected as the EMBS or IEMBS. At Wave 1, these individuals were classified as TSMs. We attempt to interview TSM participants in successive waves as long as they live in the household of an OSM.

A male TSM who fathers a child with an OSM female becomes a Permanent Sample Member (PSM). PSMs are treated in the same way as OSMs in the following rules. In sum, TSMs are not followed for interviews when they leave the household, but OSMs and PSMs are.

The following rules in the BHPS were different: TSMs of either gender became PSM and were followed in successive waves when they parented a child with another OSM. When the BHPS sample joined the UKHLS sample in Wave 2, the Understanding Society following rules started to apply.

For panel maintenance, ISER maintains a database of information on respondents so we can send communications to them and to allocate interviewers. This information is vital for minimising attrition. The data base builds on contact information collected during the survey interviews, and is updated throughout the year. There are, for example, new addresses, household splits and moves out of the country or into an institution. Change of address cards were also returned to ISER in cases where a whole household moved or a new resident returned the card giving the forwarding address. It is possible for ISER to be notified of some deaths through these means.

A between-wave-mailing is also used to help maintain contact with participants and update addresses. The mailing has a report of research findings, an address confirmation slip and materials to encourage registration with the participant website. The participant website can be seen at
https://www.understandingsociety.ac.uk/participants

2.3.4. RESPONSE OUTCOMES

This section summarises the response outcomes for Waves 1-7 in the UKHLS. For response outcomes in the BHPS see Taylor (2010), Section A4.

2.3.4.1. Wave 1

The Wave 1 fieldwork started on 8th January 2009 and ended on the 7th March 2011 (including the re-issue period). In total, interviews were achieved in 30,169 households (26,089 in the GPS, 4,080 in the EMBS), with full or proxy interviews with 50,994 individuals (43,674 in the GPS and 7,320 in the EMBS).

Table 2 and Table 3 present the household and individual response rates for Wave 1. The individual response rates are for co-operating households only.

The response rates for the EMBS component do not make any correction for the probability of non-interviewed cases being ineligible. The estimated response rate taking this factor into account is substantially higher.

Table 2: Household response rates among eligible households, Wave 1

	GPS			EMBS
	Great Britain	Northern Ireland	Total	
Full interview	57.1%	60.9%	57.3%	39.9%
Proxy interview	8.1%	11.0%	8.3%	28.0%
Refusal	33.9%	27.4%	33.6%	29.0%
Other non-interview	0.8%	0.7%	0.8%	3.1%
Total	43,267	2,107	45,374	10,077

Table 3: Individual response rates, Wave 1

	GPS			EMBS
	Great Britain	Northern Ireland	Total	
Full interview	82.0%	77.3%	81.8%	72.4%
Proxy interview	5.3%	3.5%	5.2%	6.9%
Refusal	6.5%	9.2%	6.7%	8.7%
Other non-interview	6.1%	9.9%	6.3%	12.1%
Total	47,615	2,584	50,199	9,237

2.3.4.2. Wave 2

The Wave 2 fieldwork started on 12th January 2010 and ended on the 27th March 2012 (including the re-issue period).

Household response rates for Wave 2 are shown in Table 4. The table separates the different samples. The GPS consists of respondents in Great Britain and Northern Ireland. The EMBS households are only located in Great Britain. The former-BHPS sample consists of the *Living in Britain* sample (started in 1991), the *Living in Scotland* and *Living in Wales* boost samples (started in 1999) and the *Northern Ireland Household Panel Survey* (NIHPS, started in 2001), also a boost sample.

Ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals or emigrated from the UK. For the former-BHPS component, ineligible households would also include households which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

Fully responding households are those in which the household is successfully enumerated, the household questionnaire is completed, and all eligible adults give an individual interview. Partially responding households are those where the household is enumerated and a household questionnaire is done, and at least one eligible adult but not all eligible adults complete an individual interview.

Table 4: Household response rates, Wave 2

	UKHLS GPS		EMBS	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	16,003 61.8%	873 65.3%	2,030 49.3%	3,112 66.5%	793 64.3%	833 64.1%	990 73.4%	24,634 61.7%
Partially responding	3,888 15.0%	221 16.5%	749 18.2%	504 10.8%	114 9.2%	165 12.7%	153 11.4%	5,794 14.5%
All responding	19,891 76.8%	1,094 81.9%	2,779 67.5%	3,616 77.2%	907 73.5%	998 76.8%	1,143 84.8%	30,428 76.2%
Non-contact	1,116 4.3%	22 1.7%	299 7.3%	217 4.6%	73 5.9%	62 4.8%	33 2.5%	1,825 4.6%
Untraced mover	1,450 5.6%	50 3.7%	411 10.0%	181 3.9%	49 4.0%	50 3.9%	43 3.2%	2,235 5.6%
Refusal	3,359 13.0%	162 12.1%	600 14.6%	648 13.8%	199 16.1%	185 14.2%	117 8.7%	5,281 13.2%
Other non-interview	94 0.4%	8 0.6%	28 0.7%	20 0.4%	6 0.5%	5 0.4%	12 0.9%	173 0.4%
Total*	25,910	1,336	4,117	4,682	1,234	1,300	1,348	39,942

* Base is all households issued to the field for Wave 2, minus any found to have become ineligible.

Household response rates were higher in Northern Ireland than in the rest of the UK. The household response rate for the continuing *Understanding Society* GPS was 76.8% in GB and 81.9% in Northern Ireland. The household response rates for the former-BHPS component were similar to the *Understanding Society* GPS. Among the samples in GB, the *Living in Britain* households had the highest response rate at 77.2%. The *Living in Wales* households had a similar response rate to the *Living in Britain* sample (76.8%), whilst *Living in Scotland* had a lower response rate at 73.5%. The NIHPS had a higher household response rate than in GB, with 84.8%. The response rates for the BHPS samples in GB were disappointing, given that this was, in effect, Wave 19 for many households. However, the lower response rate may have been due to the change in the fieldwork agency, interviewers, survey name, and logo. Interestingly, in Northern Ireland where the survey name and logo changed, but the fieldwork agency and so the interviewer stayed the same as in NIHPS, the response rate was much higher.

Non-contact rates were lower in Northern Ireland than in GB. The level of untraced movers was higher for the *Understanding Society* GPS in GB than in the former-BHPS. The levels of non-contact and untraced movers were highest in the EMBSs, possibly reflecting their younger average age, concentration in large urban areas, and higher level of mobility. Within the former-BHPS, the level of untraced movers was higher than in the past. This is likely to be due to the longer gap between waves of interview. The interviews for the former BHPS sample for Wave 2 of *Understanding Society* took place throughout 2010 and into the early months of 2011. The previous interview for most of these households was between September and December 2008. As the gap between the Wave 18 BHPS interview and the

Wave 2 *Understanding Society* interview increased, so did the level of untraced movers.

Refusals were generally higher in GB than in Northern Ireland. Refusals are expected to be higher at the second wave of a longitudinal study than at subsequent waves. The higher than expected refusal rate for the former BHPS sample, particularly those in GB, may be due to the aforementioned change in the name and logo of the Study as well as the change in fieldwork agency and thus, for most households, a change of interviewer.

Table 5 shows the Wave 2 cross-sectional response rates for adults. Where a household responded, we have an individual-level outcome for all adults. Where a household did not respond, we have assigned the household nonresponse outcome to the adults who were issued to that household. From this we can see, for example, that we were not able to interview 7,229 adults in the *Understanding Society* GPS in GB because they were residing in households that refused to participate at Wave 2. In the GB samples of the former BHPS there is a relatively small group of households who only give telephone interviews.

On a longitudinal study, such as *Understanding Society*, researchers are typically interested in having pairs of observations on the same individual to investigate individual-level change over time. Table 6 takes as the baseline all those who gave a full interview at the previous wave, and shows their outcome at Wave 2. For the former BHPS samples, the previous wave was Wave 18 (i.e., for the *Living in Britain sample*), Wave 10 (i.e., for the *Living in Scotland* and *Living in Wales samples*) or Wave 8 (i.e., for the NIHPS); all collected in 2008.

Once more, we see that there is a higher re-interview rate in the Northern Ireland samples than in GB. The lowest re-interview rate is in the EMBS, largely due to a higher level of non-contacted households or households who moved but could not be traced. Interestingly, the re-interview rate was higher in the GPS GB sample than in the three samples that made up the former-BHPS GB samples. Overall, in the Waves 1 and 2 data, pairs of observations are available for 45,836 adults. If proxy and telephone interviews are included, this increases to 47,282 adults.

For more detail, please see the working paper on nonresponse and attrition (Lynn, Burton et al. 2012).

Table 5: Wave 2 cross-sectional individual adult response rates by sample origin

	UKHLS GPS		**EMBS**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	32,381 60.8%	1,770 62.3%	4,978 46.3%	6,140 61.6%	1,461 58.1%	1,651 59.6%	2,008 71.7%	50,389 59.4%
Proxy interview	2,722 5.1%	87 3.1%	615 5.7%	253 2.5%	49 2.0%	86 3.1%	58 2.1%	3,870 4.6%
Telephone interview	--	--	--	202 2.0%	66 2.6%	58 2.1%	--	326 0.4%
Other non-interview	1,184 2.2%	126 4.4%	472 4.4%	200 2.0%	58 2.3%	64 2.3%	107 3.8%	2,211 2.6%
Refusal	2,104 4.0%	218 7.7%	511 4.8%	341 3.4%	92 3.7%	114 4.1%	133 4.8%	3,513 4.1%
Household non-contact	3,338 6.3%	125 4.4%	1,156 10.8%	555 5.6%	210 8.3%	155 5.6%	111 4.0%	5,650 6.7%
Household refusal	7,229 13.6%	350 12.3%	1,743 16.2%	1,493 15.0%	400 15.9%	427 15.4%	203 7.2%	11,845 14.0%
Household other non-interview	118 0.2%	5 0.2%	60 0.6%	29 0.3%	9 0.4%	6 0.2%	4 0.1%	231 0.3%
Household untraced	4,178 7.9%	159 5.6%	1,207 11.2%	754 7.6%	172 6.8%	208 7.5%	178 6.4%	6,856 8.1%
Total	53,254	2,840	10,742	9,967	2,517	2,769	2,802	84,891

Table 6: Wave 2 longitudinal re-interview rates for adults with full interview at Wave 1 by sample origin

	UKHLS GPS		**EMBS**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	29,646 74.3%	1,640 81.0%	4,200 62.2%	5,633 69.4%	1,335 67.8%	1,507 67.6%	1,875 83.3%	45,836 72.4%
Proxy interview	775 1.9%	16 0.8%	188 2.8%	97 1.2%	17 0.9%	38 1.7%	15 0.7%	1,146 1.8%
Telephone interview	--	--	--	184 2.3%	59 3.0%	57 2.6%	--	300 0.5%
Other non-interview	334 0.8%	22 1.1%	157 2.3%	73 0.9%	16 0.8%	28 1.3%	32 1.4%	662 1.1%
Refusal	316 0.8%	25 1.2%	94 1.4%	53 0.7%	13 0.7%	15 0.7%	11 0.5%	527 0.8%
Household non-contact	1,890 4.7%	68 3.4%	500 7.4%	376 4.6%	126 6.4%	96 4.3%	34 1.5%	3,092 4.9%
Household refusal	4,144 10.4%	167 8.3%	734 10.9%	965 11.9%	245 12.4%	260 11.7%	109 4.8%	6,633 10.5%
Household other non-interview	65 0.2%	2 0.1%	31 0.5%	18 0.2%	3 0.2%	2 0.1%	2 0.1%	123 0.2%
Household untraced	2,252 5.6%	63 3.1%	639 9.5%	439 5.4%	105 5.3%	140 6.3%	105 4.7%	3,744 5.9%
Household ineligible	507 1.3%	22 1.1%	208 3.1%	280 3.5%	51 2.6%	86 3.9%	68 3.0%	1,222 1.9%
Total	39,929	2,025	6,751	8,118	1,970	2,229	2,251	63,285

2.3.4.3. Wave 3

The Wave 3 fieldwork started on 7th January 2011 and ended on the 12th July 2013 (including the re-issue period). Table 7 shows the household response rates for Wave 3 of *Understanding Society*. The table separates the different samples as in Table 4. As before, ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals, emigrated from the UK or which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

Table 7: Household response rates, Wave 3

	UKHLS GPS		**EMBS**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	13,629 57.1%	680 55.3%	1,566 42.9%	2,762 66.2%	664 62.2%	726 62.2%	833 66.4%	20,860 57.3%
Partially responding	4,341 18.2%	293 23.8%	951 26.0%	653 15.7%	153 14.3%	218 18.7%	246 19.6%	6,855 18.8%
All responding	17,970 75.3%	973 79.1%	2,517 68.9%	3,415 81.8%	817 76.5%	944 80.9%	1,079 86.0%	27,715 76.1%
Non-contact	1,056 4.4%	91 7.4%	277 7.6%	186 4.5%	77 7.2%	55 4.7%	54 4.3%	1,796 4.9%
Untraced mover	1,458 6.1%	44 3.6%	351 9.6%	133 3.2%	48 4.5%	50 4.3%	51 4.1%	2,135 5.9%
Refusal	3,088 12.9%	102 8.3%	461 12.6%	395 9.5%	115 10.8%	105 9.0%	55 4.4%	4,321 11.9%
Other non-interview	295 1.2%	19 1.6%	47 1.3%	44 1.1%	11 1.0%	13 1.1%	15 1.2%	444 1.2%
Total*	23,867	1,229	3,653	4,173	1,068	1,167	1,254	36,411
* Base is all households issued to the field for Wave 3, minus any found to have become ineligible.								

Household response rates (including partial household response) were higher in Northern Ireland than in the rest of the UK, as at Wave 2. The household response rate for the continuing *Understanding Society* GPS was 75.3% in GB, and 79.1% in Northern Ireland. The household response rates for the former-BHPS samples were higher than the *Understanding Society* GPS, both in terms of overall household rates and fully responding households, although there is only a slight difference between the overall household rate for *Living in Scotland* and the British GPS rate. Among the samples in GB, the *Living in Britain* households – who have been part of the sample for the longest time – had the highest response rate at 81.9%. The *Living in Wales* households had a similar response rate to the *Living in Britain* sample (80.9%), whilst *Living in Scotland* had a lower response rate at 76.5%. The NIHPS had a higher household response rate than the GB samples, with 86%.

The EMBS had the lowest household response rate of all the samples in the Study. The main reasons for the lower response appear to be a higher level of non-contacts and untraced movers in this sample. The EMBS is concentrated in areas of high ethnic minority density, which tend to be very urbanized areas, particularly in London. Residential mobility is higher among those living in these types of area,

which may contribute to the higher levels of untraced movers. Non-contact rates also tend to be higher in cities, particularly in areas of tower blocks, flats or apartments with a common – often locked – entrance. The refusal rates for the EMBS are similar to those of the GPS in Britain and only a couple of percentage points higher than the *Living in Scotland* sample.

Refusal rates were still higher for the former-BHPS samples in Britain than they had been in the later years of the BHPS. The refusal rate from Waves 14-18 were around 6%, compared to 9-11% at Wave 3 of *Understanding Society*. This may reflect some resistance after the change in fieldwork agency.

Table 8 shows the cross-sectional response rates for adults in Wave 3. Where a household responded, we have an individual-level outcome for all adults. Where a household did not respond, we have assigned the household nonresponse outcome to the adults who were issued to that household. At Wave 3, the telephone interview was conducted using the same instrument as the face-to-face instruments but with slight changes to reflect the aural rather than visual delivery of the questionnaire (e.g., there were no showcards). Apart from these changes, the content of the telephone questionnaire was the same as the face-to-face questionnaire and so these interviews are classified as “full interviews” below.

Table 9 takes as the baseline all those who gave a full interview at the previous wave, and shows their outcome at Wave 3. Once more, we see that there is a higher re-interview rate in the NI samples than in GB. The lowest re-interview rate is in the EMBS, largely due to a higher level of refusal households or households who moved but could not be traced. Unlike at Wave 2, the re-interview rate was lower in the GPS GB than in the three samples that made up the former-BHPS GB samples. The lower re-interview rate at Wave 2 for these BHPS sample households may reflect a “blip” caused by the changes in fieldwork agency, the name and branding of the survey. Overall, in the Waves 2 and 3 data, pairs of observations are available for 45,903 adults. If proxy interviews are included, this increases to 49,708 adults.

Table 8: Wave 3 cross-sectional individual adult response rates by sample origin

	UKHLS GPS		**EMBS**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	29,135 61.3%	1,550 59.5%	4,432 47.8%	5,959 70.4%	1,377 66.0%	1,604 67.6%	1,846 72.5%	45,903 61.3%
Proxy interview	2,516 5.3%	47 1.8%	667 7.2%	333 3.9%	65 3.1%	107 4.5%	70 2.8%	3,805 5.1%
Other non-interview	1,039 2.2%	150 5.8%	426 4.6%	144 1.7%	54 2.6%	60 2.5%	107 4.2%	1,980 2.6%
Refusal	2,051 4.3%	207 7.9%	490 5.3%	322 3.8%	68 3.3%	111 4.7%	142 5.6%	3,391 4.5%
Household non-contact	2,274 4.8%	253 9.7%	834 9.0%	418 4.9%	176 8.4%	135 5.7%	134 5.3%	4,224 5.6%
Household refusal	7,826 16.5%	295 11.3%	1,611 17.4%	1,005 11.9%	260 12.4%	255 10.7%	149 5.9%	11,401 15.2%
Household other non-interview	414 0.9%	34 1.3%	107 1.2%	67 0.8%	15 0.7%	22 0.9%	25 1.0%	684 0.9%
Household untraced	2,262 4.8%	71 2.7%	698 7.5%	219 2.6%	71 3.4%	80 3.4%	73 2.9%	3,474 4.6%
Total	47,517	2,607	9,265	8,467	2,086	2,374	2,546	74,862

Table 9: Wave 3 longitudinal re-interview rates for adults with full interview at Wave 2 by sample origin

	UKHLS GPS		EMBS	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	25,853 78.5%	1,440 80.5%	3,519 69.4%	5,424 83.8%	1,249 80.6%	1,434 82.3%	1,747 85.3%	40,666 78.8%
Proxy interview	706 2.1%	16 0.9%	201 4.0%	99 1.5%	22 1.4%	37 2.1%	34 1.7%	1,115 2.2%
Other non-interview	295 0.9%	48 2.7%	109 2.2%	56 0.9%	14 0.9%	22 1.3%	27 1.3%	571 1.1%
Refusal	287 0.9%	30 1.7%	77 1.5%	47 0.7%	10 0.7%	17 1.0%	15 0.7%	483 1.0%
Household non-contact	832 2.5%	44 2.5%	212 4.2%	158 2.4%	57 3.7%	51 2.9%	34 1.7%	1,388 2.7%
Household refusal	2,919 8.9%	108 6.0%	418 8.2%	312 4.8%	99 6.4%	82 4.7%	59 2.9%	3,997 7.8%
Household other non-interview	258 0.8%	21 1.2%	40 0.8%	42 0.7%	11 0.7%	15 0.9%	22 1.1%	409 0.8%
Household untraced	1,285 3.9%	52 2.9%	352 6.9%	230 3.6%	52 3.4%	54 3.1%	78 3.8%	2,103 4.1%
Household ineligible	483 1.5%	29 1.6%	142 2.8%	108 1.7%	36 2.3%	31 1.8%	33 1.6%	862 1.7%
Total	32,918	1,788	5,070	6,476	1,550	1,743	2,049	51,594

2.3.4.4. Wave 4

The Wave 4 fieldwork started on 8h January 2012 and ended on the 19th June 2013 (including the re-issue period). Table 10 shows the household response rates for Wave 4, separated by samples. As before, ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals, emigrated from the UK or which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

The fully-responding household response rates for the former-BHPS samples were higher than the *Understanding Society* samples. The household response rate was lower in the former *Living in Scotland* than the other BHPS samples.

The EMBS continues to have the lowest household response rate of all the samples in the Study. The main reason for the lower response continues to be a higher level of non-contacts and untraced movers in this sample. The refusal rates for the EMBS at Wave 4 are slightly higher than those of the GPS in Britain and only a couple of percentage points higher than the *Living in Scotland* sample.

Table 11 shows the cross-sectional response rates for adults in Wave 4. Where a household responded, we have an individual-level outcome for all adults. Where a household did not respond, we have assigned the household nonresponse outcome to the adults who were issued to that household.

Table 12 shows the individual re-interview rates, by sample. The re-interview rate at Wave 4 is higher than it was at Wave 3; 82.7% of those who gave a full interview at Wave 3 also gave one at Wave 4, compared to a 78.8% re-interview rate between Waves 2 and 3. In the UKHLS GPS, the re-interview rate is higher for the British sample than the Northern Ireland sample. The opposite is the case for the former-BHPS samples; the Northern Ireland sample has the highest re-interview rate.

Table 10: Household response rates, Wave 4

	UKHLS GPS		EMBS	Former-BHPS				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	12,897 62.0%	603 56.9%	1,505 46.9%	2,579 68.9%	594 63.2%	666 62.7%	789 67.9%	19,633 61.4%
Partially responding	3,958 19.0%	233 22.0%	819 25.5%	600 16.0%	146 15.5%	216 20.3%	209 18.0%	6,181 19.3%
All responding	16,855 81.0%	836 78.9%	2,324 72.4%	3,179 84.9%	740 78.7%	882 83.0%	998 85.9%	25,814 80.7%
Non-contact	858 4.1%	63 5.9%	244 7.6%	113 3.0%	41 4.4%	39 3.7%	45 3.9%	1,403 4.4%
Untraced mover	754 3.6%	50 4.7%	200 6.2%	87 2.3%	33 3.5%	39 3.7%	46 4.0%	1,209 3.8%
Refusal	2,134 10.3%	96 9.1%	392 12.2%	325 8.7%	100 10.6%	94 8.9%	53 4.6%	3,194 10.0%
Other non-interview	202 1.0%	15 1.4%	50 1.6%	38 1.0%	26 2.8%	8 0.8%	20 1.7%	359 1.1%
Total*	20,803	1,060	3,210	3,742	940	1,062	1,162	31,979
* Base is all households issued to the field for Wave 4, minus any found to have become ineligible.								

Table 11: Wave 4 cross-sectional individual adult response rates by sample origin

	UKHLS GPS		**EMBS**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	27,643 67.0%	1,341 60.7%	4,236 52.0%	5,547 73.4%	1,222 66.9%	1,479 68.7%	1,749 74.1%	43,217 66.0%
Proxy interview	2,654 6.4%	36 1.6%	666 8.2%	330 4.4%	74 4.1%	115 5.3%	40 1.7%	3,915 6.0%
Other non-interview	740 1.8%	113 5.1%	290 3.6%	104 1.4%	53 2.9%	52 2.4%	88 3.7%	1,439 2.2%
Refusal	1,676 4.1%	180 8.2%	394 4.8%	298 3.9%	62 3.4%	115 5.3%	153 6.5%	2,878 4.4%
Household non-contact	1,958 4.8%	150 6.8%	756 9.3%	243 3.2%	84 4.6%	95 4.4%	110 4.7%	3,396 5.2%
Household refusal	5,209 12.6%	281 12.7%	1,328 16.3%	842 11.1%	243 13.3%	224 10.4%	139 5.9%	8,266 12.6%
Household other non-interview	296 0.7%	35 1.6%	127 1.6%	52 0.7%	40 2.2%	17 0.8%	27 1.1%	594 0.9%
Household untraced	1,076 2.6%	72 3.3%	346 4.3%	143 1.9%	50 2.7%	56 2.6%	56 2.4%	1,799 2.8%
Total	41,252	2,208	8,143	7,559	1,828	2,153	2,362	65,505

Table 12: Wave 4 longitudinal re-interview rates for adults with full interview at Wave 3 by sample origin

	UKHLS GPS		**EMBS**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	24,705 83.7%	1,246 79.5%	3,375 74.6%	5,070 84.3%	1,095 78.7%	1,351 82.2%	1,641 87.5%	38,483 82.7%
Proxy interview	609 2.1%	10 0.6%	137 3.0%	101 1.7%	26 1.9%	30 1.8%	7 0.4%	920 2.0%
Other non-interview	198 0.7%	25 1.6%	86 1.9%	30 0.5%	18 1.3%	11 0.7%	24 1.3%	392 0.8%
Refusal	183 0.6%	18 1.2%	62 1.4%	29 0.5%	12 0.9%	8 0.5%	12 0.6%	324 0.7%
Household non-contact	663 2.3%	60 3.8%	201 4.4%	79 1.3%	25 1.8%	26 1.6%	41 2.2%	1,095 2.4%
Household refusal	1,838 6.2%	131 8.4%	384 8.5%	385 6.4%	116 8.3%	106 6.5%	57 3.0%	3,017 6.5%
Household other non-interview	164 0.6%	15 1.0%	57 1.3%	32 0.5%	30 2.2%	7 0.4%	21 1.1%	326 0.7%
Household untraced	857 2.9%	40 2.6%	167 3.7%	221 3.7%	50 3.6%	88 5.4%	41 2.2%	1,464 3.1%
Household ineligible	315 1.1%	23 1.5%	58 1.3%	69 1.2%	19 1.4%	17 1.0%	32 1.7%	533 1.1%
Total	29,532	1,568	6,016	1,391	1,644	1,876	4,527	46,554

2.3.4.5. Wave 5

The Wave 5 fieldwork started on 8th January 2013 and ended on the 5th June 2015 (including the re-issue period). Table 13 shows the household response rates for Wave 5 separated by samples. As before, ineligible households have been removed from the table, these would include households where all sample members had died, consist of only TSM individuals, emigrated from the UK or which have merged with a previous wave household (for example, an adult moving back to live with his or her parents who are also part of the sample).

Table 13: Household response rates, Wave 5

	UKHLS GPS		**EMBS**	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	12,357 65.3%	552 59.4%	1,412 50.0%	2,515 72.0%	572 68.3%	646 65.5%	696 65.7%	18,750 64.6%
Partially responding	3,543 18.7%	214 23.0%	748 26.5%	533 15.3%	121 14.5%	193 19.6%	211 19.9%	5,563 19.2%
All responding	15,900 84.1%	766 82.4%	2,160 76.5%	3,048 87.2%	693 82.8%	839 85.1%	907 85.7%	24,313 83.8%
Non-contact	638 3.4%	56 6.0%	162 5.7%	128 3.7%	42 5.0%	53 5.4%	57 5.4%	1,136 3.9%
Untraced mover	455 2.4%	30 3.2%	126 4.5%	45 1.3%	10 1.2%	15 1.5%	30 2.8%	711 2.5%
Refusal	1,746 9.2%	64 6.9%	345 12.2%	238 6.8%	78 9.3%	70 7.1%	55 5.2%	2,596 8.9%
Other non-interview	177 0.9%	14 1.5%	30 1.1%	36 1.0%	14 1.7%	9 0.9%	10 0.9%	290 1.0%
Total*	18,916	930	2,823	3,495	837	986	1,059	29,046
* Base is all households issued to the field for Wave 5, minus any found to have become ineligible.								

The fully-responding household response rates for the former-BHPS samples were higher than the *Understanding Society* samples. The household response rate was lower in the former *Living in Scotland* than the other BHPS samples.

The EMBS continues to have the lowest household response rate of all the samples in the Study. The main reason for the lower response continues to be a higher level of non-contacts and untraced movers in this sample, along with a refusal rate which is almost three percentage points higher than the next highest rate.

Table 14 shows the cross-sectional response rates for adults in Wave 5. Where a household responded, we have an individual-level outcome for all adults. Where a household did not respond, we have assigned the household nonresponse outcome to the adults who were issued to that household.

Table 14: Wave 5 cross-sectional individual adult response rates by sample origin

	UKHLS GPS		EMBS	Former-BHPS				Total
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Full interview	26,318 67.0%	1,215 60.6%	3,961 51.5%	5,386 73.9%	1,167 68.2%	1,434 69.2%	1,560 68.5%	41,041 65.9%
Proxy interview	2,531 6.5%	39 2.0%	748 9.7%	319 4.4%	68 4.0%	88 4.3%	47 2.1%	3,840 6.2%
Other non-interview	575 1.5%	105 5.2%	204 2.7%	87 1.2%	34 2.0%	39 1.9%	98 4.3%	1,142 1.8%
Refusal	1,496 3.8%	157 7.8%	322 4.2%	254 3.5%	54 3.2%	116 5.6%	138 6.1%	2,537 4.1%
Household non-contact	2,911 7.4%	217 10.8%	920 12.0%	475 6.5%	144 8.4%	177 8.5%	195 8.6%	5,039 8.1%
Household refusal	4,323 11.0%	180 9.0%	1,222 15.9%	606 8.3%	198 11.6%	176 8.5%	161 7.1%	6,866 11.0%
Household other non-interview	472 1.2%	29 1.5%	88 1.1%	89 1.2%	34 2.0%	21 1.0%	28 1.2%	761 1.2%
Household untraced	640 1.6%	63 3.1%	222 2.9%	72 1.0%	13 0.8%	22 1.1%	49 2.2%	1,081 1.7%
Total	39,266	2,005	7,687	7,288	1,712	2,073	2,276	62,307

Table 15: Wave 5 longitudinal re-interview rates for adults with full interview at Wave 4 by sample origin

	UKHLS GPS		EMBS	Former-BHPS				Total
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Full interview	23,984 85.8%	1,127 83.4%	3,349 78.6%	4,918 87.5%	1,044 84.5%	1,297 87.1%	1,472 83.5%	37,191 85.2%
Proxy interview	521 1.9%	14 1.0%	165 3.9%	85 1.5%	17 1.4%	12 0.8%	27 1.5%	841 1.9%
Other non-interview	144 0.5%	25 1.9%	60 1.4%	21 0.4%	8 0.7%	11 0.7%	22 1.3%	291 0.7%
Refusal	176 0.6%	12 0.9%	58 1.4%	33 0.6%	4 0.3%	10 0.7%	21 1.2%	314 0.7%
Household non-contact	508 1.8%	63 4.7%	135 3.2%	88 1.6%	27 2.2%	35 2.4%	65 3.7%	921 2.1%
Household refusal	1,397 5.0%	55 4.1%	310 7.3%	209 3.7%	80 6.5%	56 3.8%	74 4.2%	2,181 5.0%
Household other non-interview	143 0.5%	15 1.1%	29 0.7%	35 0.6%	12 1.0%	11 0.7%	11 0.6%	256 0.6%
Household untraced	792 2.8%	27 2.0%	114 2.7%	168 3.0%	26 2.1%	41 2.8%	45 2.6%	1,213 2.8%
Household ineligible	288 1.0%	13 1.0%	43 1.0%	64 1.1%	17 1.4%	16 1.1%	27 1.5%	468 1.0%
Total	27,953	1,351	4,263	5,621	1,235	1,489	1,764	43,676

Table 15 shows the individual re-interview rates, by sample. The re-interview rate at Wave 5 is higher than it was at Wave 4; 85.2% of those who gave a full interview at Wave 4 also gave one at Wave 5, compared to a re-interview rate of 82.7% between Waves 3 and 4, and a re-interview rate of 78.8% between Waves 2 and 3. The highest re-interview rates are for the former BHPS sample (particularly in the original 1991 *Living in Britain* and the 2001 NIHPS samples) and the GPS. The lowest re-interview rate is once more in the ethnic minority boost, with a higher household refusal and non-contact rate. The re-interview rate in Northern Ireland is almost the same for the 2009 GPS and the 2001 NIHPS samples.

2.3.4.6. Wave 6

The Wave 6 fieldwork started on 8th January 2014 and ended on the 11th May 2016 (including the re-issue period). Table 16 shows the household response rates for Wave 6 of *Understanding Society*. The table separates the different samples as in Table 7, above.

The response at Wave 6 shows a fall compared to Wave 5. This may have been due to the change in the fieldwork agencies, which meant that almost all households had a new interviewer. This is particularly so in Northern Ireland (NI), which had a much higher refusal rate for both the Northern Ireland component of the GPS and the former Northern Ireland Household Panel Survey.

Tables 16 and 17 also include the new IEMBS which was issued during Year 2 of Wave 6. The slight decreases in response rates are reflected in the cross-sectional response and re-interview rates. Note that inclusion of the new IEMBS means the "Total" column is not comparable to previous waves.

Table 17 shows the cross-sectional response rates for adults in Wave 6, and Table 18 shows the individual re-interview rates, by sample.

Table 16: Household response rates, Wave 6

	UKHLS GPS		EMBS	IEMBS	Former-BHPS				Total
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Fully responding	11,258 60.9%	534 55.0%	1,239 44.7%	1,675 18.1%	2,405 70.1%	561 69.5%	580 60.0%	644 59.1%	18,896 50.0%
Partially responding	3,216 17.4%	170 17.5%	653 23.6%	1,013 10.9%	486 14.2%	106 13.1%	182 18.8%	189 17.3%	6,015
All responding	14,474 78.3%	704 72.5%	1,892 68.3%	2,688 29.0%	2,891 84.3%	667 82.6%	762 78.8%	833 76.4%	24,911 65.9%
Non-contact	839 4.5%	39 4.0%	234 8.4%	3,478 37.5%	107 3.1%	35 4.3%	50 5.2%	38 3.5%	4,820 12.8%
Untraced mover	1,186 6.4%	57 5.9%	274 9.9%	- -	146 4.3%	32 4.0%	43 4.5%	66 6.1%	1,804 4.8%
Refusal	1,644 8.9%	133 13.7%	298 10.8%	2,559 27.6%	224 6.5%	61 7.6%	73 7.6%	123 11.3%	5,115 13.5%
Other non-interview	355 1.9%	38 3.9%	74 2.7%	542 5.9%	65 1.9%	12 1.5%	39 4.0%	30 2.8%	1,155 3.1%
Total*	18,498	971	2,772	9,267	3,433	807	967	1,090	37,805
* Base is all households issued to the field for Wave 6, minus any found to have become ineligible.									

Table 17: Wave 6 cross-sectional individual adult response rates by sample origin

	UKHLS GPS		EMBS	IEMBS*	Former-BHPS				Total
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Full interview	23,919 65.0%	1,145 57.6%	3,549 49.8%	4,458 69.5%	5,100 74.0%	1,129 71.4%	1,282 65.5%	1,439 63.1%	42,021 64.6%
Proxy interview	1,985 5.4%	71 3.6%	591 8.3%	196 3.1%	254 3.7%	42 2.6%	96 4.9%	87 3.8%	3,322 5.1%
Other non-interview	614 1.7%	19 1.0%	235 3.3%	527 8.2%	77 1.1%	29 1.8%	41 2.1%	42 1.8%	1,584 2.4%
Refusal	1,574 4.3%	135 6.8%	325 4.6%	653 10.2%	270 3.9%	64 4.1%	92 4.7%	138 6.1%	3,251 5.0%
Household non-contact	1,942 5.3%	89 4.5%	672 9.4%	- -	242 3.5%	79 5.0%	117 6.0%	100 4.4%	3,241 5.0%
Household refusal	4,170 11.3%	365 18.4%	1,030 14.5%	- -	607 8.8%	156 9.9%	187 9.6%	342 15.0%	6,857 10.5%
Household other non-interview	694 1.9%	78 3.9%	221 3.1%	583 9.1%	120 1.7%	33 2.1%	84 4.3%	52 2.3%	1,865 2.9%
Household untraced	1,879 5.1%	85 4.3%	507 7.1%	- -	227 3.3%	50 3.2%	59 3.0%	82 3.6%	2,889 4.4%
Total	36,777	1,987	7,130	6,418	6,897	1,582	1,958	2,282	65,031

* Note: This table is based on issued individuals. For the IEMBS there were no issued individuals, just addresses. The cases in “Household other non-interview” are those in which the enumeration grid was completed but no adult interviews were conducted

.

Table 18: Wave 6 longitudinal re-interview rates for adults with full interview at Wave 5 by sample origin

	UKHLS GPS		EMBS	Former-BHPS				Total
	UKHLS – GB	UKHLS – NI		Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Full interview	21,799 81.5%	953 77.8%	2,958 72.9%	4,705 84.9%	1,025 86.4%	1,170 79.5%	1,232 78.1%	33,842 81.0%
Proxy interview	422 1.6%	24 2.0%	163 4.0%	69 1.2%	8 0.7%	24 1.6%	26 1.7%	736 1.8%
Other non-interview	206 0.8%	6 0.5%	97 2.4%	18 0.3%	12 1.0%	14 1.0%	14 0.9%	367 0.9%
Refusal	207 0.8%	21 1.7%	58 1.4%	38 0.7%	11 0.9%	12 0.8%	20 1.3%	367 0.9%
Household non-contact	721 2.7%	26 2.1%	193 4.8%	90 1.6%	28 2.4%	46 3.1%	43 2.7%	1,147 2.7%
Household refusal	1,716 6.4%	120 9.8%	291 7.2%	224 4.0%	41 3.5%	74 5.0%	151 9.6%	2,617 6.3%
Household other non-interview	402 1.5%	27 2.2%	75 1.9%	71 1.3%	17 1.4%	48 3.3%	29 1.8%	669 1.6%
Household untraced	931 3.5%	35 2.9%	178 4.4%	235 4.2%	31 2.6%	65 4.4%	49 3.1%	1,524 3.7%
Household ineligible	331 1.2%	13 1.1%	43 1.1%	93 1.7%	13 1.1%	19 1.3%	14 0.9%	526 1.3%
Total	26,735	1,225	4,056	5,543	1,186	1,472	1,578	41,795

2.3.4.7. Wave 7

The Wave 7 fieldwork started on 14th January 2015 and ended on 15th May 2017 (including the re-issue period). At Wave 7, the web interview was conducted using the same instrument as the face-to-face instruments but with slight changes to reflect that there is no interviewer. Apart from these changes, the content of the web questionnaire was the same as the face-to-face questionnaire and so these interviews are classified as "full interviews" below. Table 19 shows the household response rates for Wave 7 of *Understanding Society*. The table separates the different samples as in Table 7, above. Note that the Wave 7 tables exclude the dormant sample (as in previous waves). However, at Wave 7, the dormant sample for whom we had a postal address were invited to participate online, and an additional 290 adults followed the invitation (GPS-GB: 219; GPS-NI: 2; EMB: 42; LIB: 19; LIS: 4; LIW: 1; NIHPS: 3).

The response at Wave 7 shows a rise compared to Wave 6, apart from the IEMB sample. This is the second wave that the IEMB sample members have taken part and so we would expect the response rate to be lower than the rest of the sample, similar to the Wave 2 rates for the original sample. There had been a dip in the response rates at Wave 6, at the same time a new fieldwork agency won the contract, and so at that wave all sample members were contacted by a different interviewer. The higher response rates at Wave 7 may reflect greater interviewer continuity.

Table 20 shows the cross-sectional response rates for adults in Wave 7, and Table 21 shows the individual re-interview rates, by sample.

Table 19: Household response rates, Wave 7

	UKHLS GPS		EMBS	IEMBS	Former-BHPS				
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Fully responding	10,647 65.1%	510 59.9%	1,157 50.0%	1,302 45.3%	2,290 71.5%	508 68.5%	541 61.5%	601 62.1%	17,556 62.3%
Partially responding	3,089 18.9%	169 19.9%	597 25.8	570 19.8%	509 15.9%	122 16.4%	180 20.5%	185 19.1%	5,421 19.2%
All responding	13,736 84.0%	679 79.8%	1,754 75.7%	1,872 65.1%	2,799 87.4%	630 84.9%	721 81.9%	786 81.2%	22,977 81.5%
Non-contact	778 4.8%	41 4.8%	188 8.1%	382 13.3%	119 3.7%	31 4.2%	53 6.0%	44 4.6%	1,636 5.8%
Untraced mover	542 3.3%	18 2.1%	120 5.2%	298 10.4%	72 2.2%	21 2.8%	35 4.0%	28 2.9%	1,133 4.0%
Refusal	1,016 6.2%	94 11.1%	203 8.8%	231 8.0%	179 5.6%	43 5.8%	54 6.1%	91 9.4%	1,911 6.8%
Other non-interview	289 1.8%	19 2.2%	51 2.2%	92 3.2%	33 1.0%	17 2.3%	17 1.9%	19 2.0%	537 1.9%
Total*	16,361	851	2,316	2,875	3,201	742	880	968	28,194
* Base is all households issued to the field for Wave 7 excluding the dormant sample, minus any found to have become ineligible.									

Table 20: Wave 7 cross-sectional individual adult response rates by sample origin

	UKHLS GPS		**EMBS**	**IEMBS***	**Former-BHPS**				
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	**Total**
Full interview	22,832 69.8%	1,103 62.6%	3,337 55.5%	3,299 48.3%	4,880 76.0%	1,057 73.1%	1,202 66.8%	1,382 66.3%	39,092 66.1%
Proxy interview	1,566 4.8%	106 6.0%	456 7.6%	176 2.6%	262 4.1%	51 3.5%	96 5.3%	120 5.8%	2,834 4.8%
Other non-interview	637 2.0%	11 0.6%	213 3.5%	320 4.7%	77 1.2%	18 1.2%	33 1.8%	18 0.9%	1,327 2.3%
Refusal	1,706 5.2%	98 5.6%	335 5.6%	364 5.3%	285 4.4%	86 5.9%	98 5.4%	106 5.1%	3,078 5.2%
Household non-contact	1,897 5.8%	101 5.7%	605 10.1%	1,062 15.5%	283 4.4%	70 4.8%	145 8.1%	117 5.6%	4,280 7.2%
Household refusal	2,561 7.8%	274 15.5%	678 11.3%	675 9.9%	441 6.9%	101 7.0%	127 7.1%	253 12.1%	5,110 8.6%
Household other non-interview	624 1.9%	43 2.4%	143 2.4%	248 3.6%	71 1.1%	22 1.5%	23 1.3%	38 1.8%	1,212 2.1%
Household untraced	909 2.8%	27 1.5%	248 4.1%	691 10.1%	126 2.0%	42 2.9%	76 4.2%	50 2.4%	2,169 3.7%
Total	32,732	1,763	6,015	6,835	6,425	1,447	1,800	2,084	59,101

*Note: This table is based on issued individuals, excluding dormant households. For the IEMBS there were no issued individuals, just addresses. The cases in “Household other non-interview” are those in which the enumeration grid was completed but no adult interviews were conducted.

Table 21: Wave 7 longitudinal re-interview rates for adults with full interview at Wave 7 by sample origin

	UKHLS GPS		EMBS	IEMBS	Former-BHPS				
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	Total
Full interview	20,568 85.1%	961 83.7%	2,781 76.7%	2,777 61.2%	4,514 87.3%	975 86.7%	1,084 84.4%	1,203 82.9%	34,863 82.0%
Proxy interview	233 1.0%	25 2.2%	87 2.4%	55 1.2%	52 1.0%	14 1.2%	11 0.9%	35 2.4%	512 1.2%
Other non-interview	192 0.8%	3 0.3%	74 2.0%	133 2.9%	30 0.6%	7 0.6%	11 0.9%	8 0.6%	458 1.1%
Refusal	200 0.8%	12 1.1%	49 1.4%	86 1.9%	42 0.8%	13 1.2%	10 0.8%	24 1.7%	436 1.0%
Household non-contact	607 2.5%	27 2.4%	165 4.5%	468 10.3%	94 1.8%	24 2.1%	45 3.5%	35 2.4%	1,465 3.5%
Household refusal	807 3.3%	68 5.9%	165 4.6%	267 5.9%	127 2.5%	34 3.0%	37 2.9%	75 5.2%	1,580 3.7%
Household other non-interview	276 1.1%	18 1.6%	55 1.5%	125 2.8%	43 0.8%	18 1.6%	11 0.9%	12 0.9%	558 1.3%
Household untraced	738 3.1%	21 1.8%	177 4.9%	474 10.4%	171 3.3%	26 2.3%	44 3.4%	38 2.6%	1,690 4.0%
Household ineligible	550 2.3%	13 1.1%	75 2.1%	155 3.4%	100 1.9%	14 1.2%	32 2.5%	21 1.5%	960 2.3%
Total	24,171	1,148	3,628	4,540	5,173	1,125	1,285	1,451	42,521

2.3.4.8. Wave 8

The Wave 8 fieldwork started on 5th January 2016 and ended on 10th May 2018 (including the re-issue period). Table 22 shows the household response rates for Wave 8 of Understanding Society. Table 23 shows the cross-sectional response rates for adults in Wave 8, and Table 24 shows the individual re-interview rates, by sample. Note that the Wave 8 tables exclude the dormant sample (as in previous waves).

The household response at Wave 8 is a slight increase on that at Wave 7; 63.8% fully responding (62.3% at Wave 7), and 83.2% with at least one responding adult (81.5% at Wave 7). The increase was for all sample groups, except for the new IEMB sample, which saw a slight fall from 65.1% of households with at least one adult responding in Wave 7, to 61.2%. Likewise, the cross-sectional individual adult response rates increased wave-on-wave, from 66.1% of adults giving a full interview at Wave 7 to 68.2% at Wave 8. Again, this was an increase for all sample groups except the IEMB. The longitudinal re-interview rates saw a large increase, from 82.0% at Wave 7 (that is, of those who gave a full interview at Wave 6, 82% gave a full interview at Wave 7), to 86.7% at Wave 8. This increase was across all sample groups, including the IEMB.

Tables 25 and 26 split the sample by the mode in which they were issued – whether they were CAPI-first or web-first. Table 25 shows the household response rates, and Table 26 the individual longitudinal re-interview rates. It should be noted that the allocation to mode of issue was not done randomly, and so care should be taken when comparing the response by mode of issue.

Table 22: Household response rates, Wave 8

	UKHLS GPS		EMBS	IEMBS	Former-BHPS				Total
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Fully responding	10,249 66.3%	491 62.4%	1,109 53.2%	1,130 43.8%	2,242 73.2%	499 70.8%	530 64.6%	598 66.7%	16,848 63.8%
Partially responding	3,000 19.4%	167 21.2%	537 25.8%	451 17.5%	506 16.5%	121 17.2%	165 20.1%	171 19.1%	5,118 19.4%
All responding	13,249 85.7%	658 83.6%	1,646 78.9%	1,581 61.2%	2,748 89.7%	620 87.9%	695 84.7%	769 85.7%	21,966 83.2%
Non-contact	651 4.2%	32 4.1%	160 7.7%	388 15.0%	88 2.9%	31 4.4%	35 4.3%	31 3.5%	1,416 5.4%
Untraced mover	437 2.8%	17 2.2%	76 3.7%	241 9.3%	66 2.2%	13 1.8%	26 3.2%	27 3.0%	903 3.4%
Refusal	788 5.1%	69 8.8%	138 6.6%	283 11.0%	104 3.4%	26 3.7%	42 5.1%	54 6.0%	1,504 5.7%
Other non-interview	335 2.2%	11 1.4%	65 3.1%	89 3.5%	57 1.9%	15 2.1%	23 2.8%	16 1.8%	611 2.3%
Total*	15,460	787	2,085	2,582	3,063	705	821	897	26,400
* Base is all households issued to the field for Wave 8 excluding the dormant sample, minus any found to have become ineligible.									

Table 23: Wave 8 cross-sectional individual adult response rates by sample origin

	UKHLS GPS		EMBS	IEMBS*	Former-BHPS				Total
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Full interview	22,043 71.7%	1,054 66.8%	3,263 59.5%	2,860 45.7%	4,801 78.7%	1,026 75.6%	1,153 69.8%	1,365 72.7%	37,565 68.2%
Proxy interview	878 2.9%	78 4.9%	272 5.0%	158 2.5%	138 2.3%	32 2.4%	50 3.0%	76 4.1%	1,682 3.1%
Other non-interview	1,314 4.3%	48 3.0%	294 5.4%	255 4.1%	236 3.9%	48 3.5%	64 3.9%	45 2.4%	2,304 4.2%
Refusal	1,567 5.1%	87 5.5%	312 5.7%	258 4.1%	240 3.9%	70 5.2%	97 5.9%	91 4.9%	2,722 4.9%
Household non-contact	1,580 5.1%	85 5.4%	454 8.3%	1,086 17.4%	210 3.4%	76 5.6%	107 6.5%	74 3.9%	3,672 6.7%
Household refusal	1,925 6.3%	180 11.4%	535 9.8%	853 13.6%	254 4.2%	62 4.6%	100 6.1%	152 8.1%	4,061 7.4%
Household other non-interview	735 2.4%	20 1.3%	187 3.4%	225 3.6%	115 1.9%	26 1.9%	38 2.3%	39 2.1%	1,385 2.5%
Household untraced	702 2.3%	26 1.7%	169 3.1%	562 9.0%	108 1.8%	17 1.3%	42 2.5%	36 1.9%	1,662 3.0%
Total	30,744	1,578	5,486	6,257	6,102	1,357	1,651	1,878	55,053
*Note: This table is based on issued individuals, excluding dormant households.									

Table 24: Wave 8 longitudinal re-interview rates for adults with full interview at Wave 8 by sample origin

	UKHLS GPS		**EMBS**	**IEMBS**	**Former-BHPS**				**Total**
	UKHLS – GB	UKHLS – NI			Living in Britain	Living in Scotland	Living in Wales	NIHPS	
Full interview	19,955 88.3%	944 86.4%	2,688 81.9%	2,291 70.5%	4,408 91.5%	936 89.3%	1,051 88.5%	1,214 89.3%	33,487 86.7%
Proxy interview	149 0.7%	18 1.7%	52 1.6%	46 1.4%	18 0.4%	5 0.5%	6 0.5%	12 0.9%	306 0.8%
Other non-interview	455 2.0%	24 2.2%	94 2.9%	81 2.5%	87 1.8%	14 1.3%	18 1.5%	15 1.1%	788 2.0%
Refusal	154 0.7%	13 1.2%	51 1.6%	46 1.4%	20 0.4%	10 1.0%	10 0.8%	17 1.3%	321 0.8%
Household non-contact	535 2.4%	25 2.3%	124 3.8%	280 8.6%	72 1.5%	32 3.1%	31 2.6%	22 1.6%	1,121 2.9%
Household refusal	570 2.5%	47 4.3%	135 4.1%	192 5.9%	80 1.7%	24 2.3%	37 3.1%	45 3.3%	1,130 2.9%
Household other non-interview	323 1.4%	8 0.7%	48 1.5%	80 2.5%	57 1.2%	11 1.1%	13 1.1%	12 0.9%	552 1.4%
Household untraced	226 1.0%	8 0.7%	44 1.3%	157 4.8%	37 0.8%	4 0.4%	13 1.1%	12 0.9%	501 1.3%
Household ineligible	227 1.0%	6 0.6%	45 1.4%	75 2.3%	37 0.8%	12 1.2%	8 0.7%	10 0.7%	420 1.1%
Total	22,594	1,093	3,281	3,248	4,816	1,048	1,187	1,359	38,626

Table 25: Household response rates, Wave 8 by mode of issue

	CAPI-first	Web-first	Total
Fully responding	10,632 64.6%	6,216 62.6%	16,848 63.8%
Partially responding	3,233 19.6%	1,885 19.0%	5,118 19.4%
All responding	13,865 84.2%	8,101 81.6%	21,966 83.2%
Non-contact	826 5.0%	590 5.9%	1,416 5.4%
Untraced mover	533 3.2%	370 3.7%	903 3.4%
Refusal	875 5.3%	629 6.3%	1,504 5.7%
Other non-interview	371 2.3%	240 2.4%	611 2.3%
Total*	16,470	9,930	26,400
* Base is all households issued to the field for Wave 8 excluding the dormant sample, minus any found to have become ineligible.			

Table 26: Wave 8 longitudinal re-interview rates for adults with full interview at Wave 8 by mode of issue

	CAPI-first	Web-first	Total
Full interview	22,071 84.8%	11,416 90.7%	33,487 86.7%
Proxy interview	268 1.0%	38 0.3%	306 0.8%
Other non-interview	446 1.7%	342 2.7%	788 2.0%
Refusal	212 0.8%	109 0.9%	321 0.8%
Household non-contact	947 3.6%	174 1.4%	1,121 2.9%
Household refusal	957 3.7%	173 1.4%	1,130 2.9%
Household other non-interview	428 1.6%	124 1.0%	552 1.4%
Household untraced	401 1.5%	100 0.8%	501 1.3%
Household ineligible	307 1.2%	113 0.9%	420 1.1%
Total	26,037	12,589	38,626

2.4. DATA PROCESSING AND CLEANING

The data for a sample month is delivered to ISER in batches, scheduled for four months following the beginning of the fieldwork process. This interval allows time for interview re-issue, coding, and data entry from paper documents, e.g., the self-completion instruments. Data is delivered as SPSS system files, which are then exported to triple-S data exchange format and imported into a SIR database.

Quality control processes include extensive data checking to ensure that the data conform to the expected structure and to the routing and range constraints defined by the questionnaire specifications. Data anomalies are investigated to determine whether they are related to:

- the invalid specification of the questionnaire;
- the incorrect scripting of the questionnaire;
- a failure to specify that a particular constraint should be included in the questionnaire;
- an incorrect implementation of the check, or;
- a problem in exporting and/or delivering the data.

After investigation, steps may include correcting the specification, data editing, reporting the error to the fieldwork agency to be fixed in a subsequent delivery and/or a quality feedback report suggesting changes to the questionnaire or field practice in subsequent waves.

Batch-specific databases are merged into a single database, from which anonymised data is exported for the creation of public use files. Data distributions are also checked for theoretical and statistical plausibility. This checking is done through direct scrutiny and by analyses which “road-test” the data.

Last but not least, data are being routinely checked in the process of creating added-value content, such as the respondent’s age and sex, based on the information collected across all waves, or pointers to specific other members in the household such as a biological parent. See Section 3.6.

2.4.1. CODING

Understanding Society collects freetext information on respondents' job titles and the industry of the job held. Industry descriptions are coded to ONS Standard Industry Code 2007, or SIC 2007. Job titles are coded to the ONS Standard Occupational Classification 2000, or SOC 2000. Coding is undertaken using the Computer Assisted Structured Coding Tool (CASCOT) system. From Wave 3 onward job titles are coded to SOC 2010 in the first instance and look-ups are used to code to SOC 2000. We also provide SOC 1990 codes using look-ups between SOC 2000, SOC 2010 and SOC 1990.

It should be noted that there are currently some gaps in this coding for the respondents’ current and last jobs due to unavailability of look-up files in the first waves of interviews, data collection (dependent interviewing) and data processing constraints. We aim to close these gaps in the near future. Note that, following a coding exercise in Summer 2017 and 2018, the Wave 8 release data include a great deal more valid codes for ***w*_jbsoc00** (and socio-economic classifications that rely on this variable) than previous releases. Some additional valid codes could be recovered by correcting corrupted information fed forward from the BHPS, but most occupations could be coded to SOC 2000 directly from the text descriptions and comparing against valid codes assigned to similar or identical job descriptions.

Several questions, e.g. country of birth, religion, political party, national identity, and citizenship had an “other, please specify” option. These responses were coded using an automated process.

Coding was also done for an open-ended question at Waves 1, 2, and 5, which read "We've asked you a lot of questions but we also want to know what has happened in your own life that has been especially important to you. Can you please tell me anything that has happened to you, or your family, over the past year that has stood out as important?" The respondent could give up to four answers. The answers were recorded verbatim and manually coded for type of event and its subject. The question was also asked at Wave 6, but the responses have not been coded.

2.5. DOCUMENTATION OF THE SURVEY INSTRUMENTS

The text of the questionnaires in PDF format is part of the documentation provided through the UK Data Service. Questionnaires can also be found: https://www.understandingsociety.ac.uk/documentation/mainstage/questionnaires. There are household and individual questionnaires and the adult and youth self-completion instruments. The instruments 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). The CAPI instrument governs the flow of questions and recording of answers, but it is not convenient for documentation. On the Study website, we present the questionnaire in PDF format. Similar to other PDF documents, the text of the questionnaire can be searched for specific words, such as variable names or words in questions. The PDF self-completion instruments for the adult self-completion questionnaires for Waves 1and 2 as well as the youth questionnaires (all waves) correspond to the way they appeared to participants (except they have been annotated with variable names).

The principal adult questionnaires are organized in modules. Modules can be searched for in the online documentation system https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation.

Instruments and survey materials for Waves 1 to 8 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, instruments and survey materials were translated into Bengali, Gujarati, Polish, Portuguese, Punjabi (in Urdu and Gurmukhi scripts), Somali, Turkish, and Urdu.

2.5.1. READING THE QUESTIONNAIRES

Figures 1 and 2 show marked up sample pages from the questionnaire, providing information for how to interpret the questionnaire text. Note that the variable names in the questionnaire do not have the wave prefix that is applied in the data.

Figure 1 shows a marked up sample page for a question from the household interview module. Figure 2 provides an interpretation of a more complex individual level question. The question is asked about each natural or biological child, so multiple variables are associated with the question for each natural child. The variables are located in the data file **a_natchild**, which has one record for each natural child.

Figure 1: Mark-up of household questionnaire

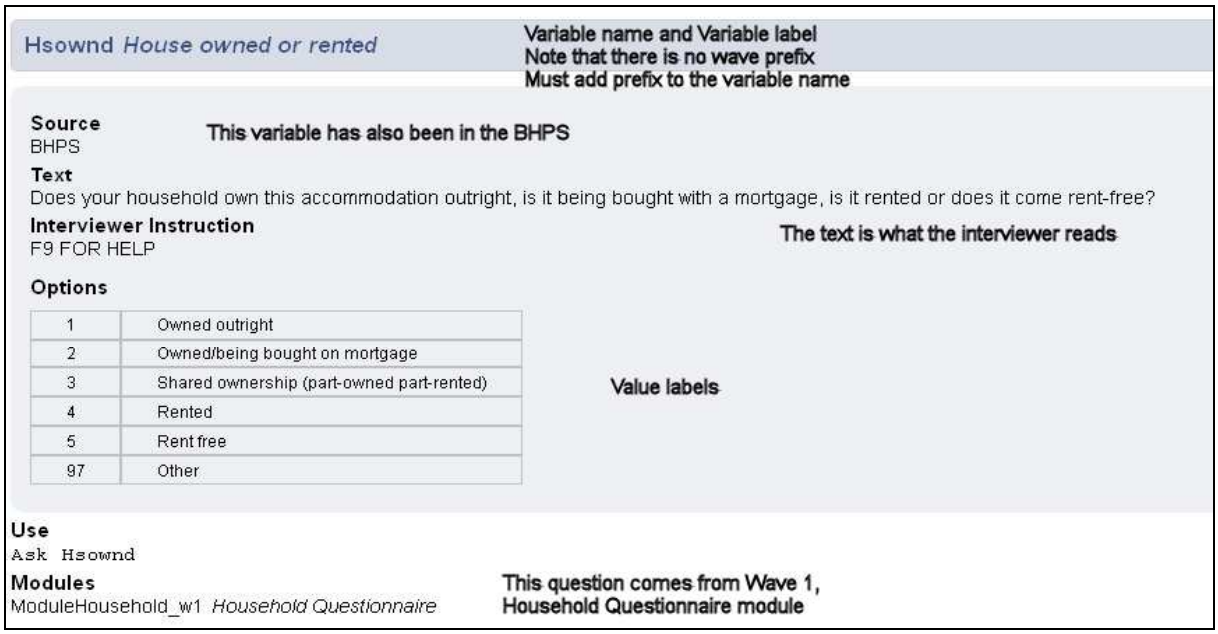
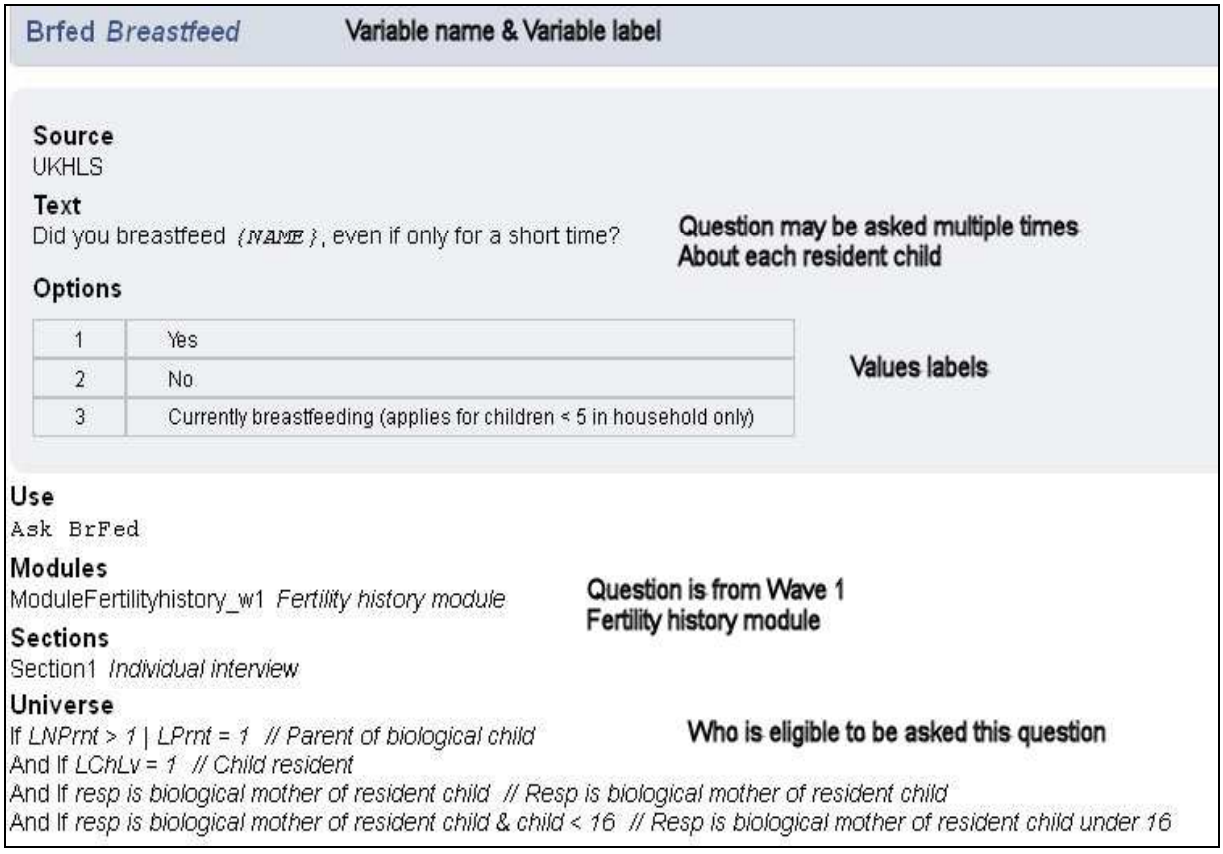


Figure 2: Mark-up of question with looping from individual questionnaire



2.5.2. SUMMARY OF QUESTIONNAIRE MODULES

The questionnaire is organised into modules. About half of the questionnaire content is collected annually, with additional modules collected at different intervals, often

every two to three years. The long-term content plan summarizes the pattern that has been collected or planned, see https://www.understandingsociety.ac.uk/documentation/mainstage/long-term-content-plan.

The long-term plan includes information about which modules are carried in the self-completion and in the youth interviews. The paper self-completion questionnaires carried at Waves 1 and 2 were not divided into modules. From Wave 3 onward, the self-completion content was carried as CASI modules, where the participant would answer the questions using the laptop.

2.5.3. CONTENT HIGHLIGHTS BY WAVE

This section provides brief overview of content covered in UKHLS Waves 1-8. For a content overview in the BHPS (and, therefore, in the *Understanding Society* harmonised BHPS files), see Taylor (2010), Sections A2-7 to A2-10. The online dataset documentation additionally provides Index Terms that assist users in identifying groups of variables on a range of different topics.

2.5.3.1. Wave 1

Wave 1 collected important baseline data. Some measures are stable, that is, not time variant. In subsequent waves, we try to collect this type of information from individuals who are new entrants to the Study. Notice that some modules are covered annually beginning in Wave 1. These represent the strongest areas for examining annual change. See for example Disability, Caring, employment related information, childcare, politics and income and benefits. These have been the focus of major longitudinal research in the BHPS and should also be a prominent focus with *Understanding Society*.

Among the Wave 1 rotating modules are: the Parents and Children module (which has content about attitudes and behaviours related to education, activities and interaction with children, and parenting practices), the Family Networks, and Environmental Behaviour module. The “Extra 5 minutes” questions included rotating modules on Remittances, Harassment, and Discrimination.

For more information on rotating modules included in the “Extra 5 minutes” questions and the waves in which these were asked see McFall, Nandi et al. (2018).

The Wave 1 self-completion questionnaire includes measures of sleep behaviour and sleep quality and a subset of items characterizing relationships with partners from the Dyadic Adjustment Scale (Spanier 1976). There are also measures of generalised trust and attitudes to risk.

2.5.3.2. Wave 2

In Wave 2 and subsequent waves, there is the Annual Event History module which is asked of persons previously interviewed. It asks about changing circumstances related to moves, marital status or cohabitation, new children including childbirth and pregnancy, new health conditions, educational experiences, and employment changes.

Wave 2 saw the introduction of a set of rotating modules related to health behaviours (nutrition, smoking, physical activity). There are modules about voluntary work and

charitable giving and important modules about savings and personal pensions. Retirement planning is an age-triggered module that is taken up again in Wave 3. Within the self-completion questionnaire, there is content on alcohol consumption, dimensions of identity and gender role attitudes.

Wave 2 also has a major module on work conditions that encompasses topics such as payment mechanisms, unions, pensions, work times, autonomy and security, and work stress.

2.5.3.3. Wave 3

There are multiple new modules for Wave 3 including those on local neighbourhoods, content on social networks in the main survey and the self-completion questionnaires, groups and organizations, use of news and media, and political self-efficacy. There is a major module on cognitive ability, see McFall (2013) for more detail about the concepts and measures of cognitive ability.

Important data related to family ties can be found in the parents and children and family networks modules, both of which are repeated from Wave 1. There is also data about child maintenance payments and relationships with children who do not live in the household. The self-completion modules have parents' reports on children including a version of the Strengths and Difficulties questionnaire, and parenting styles. The self-completion modules also include a Big 5 personality measure, sexual orientation, and several modules of questions for young adults which bring questions from the youth questionnaire into the 16-21 age group.

The youth questionnaire instrument is the same as the one used in Wave 1 (as will be the case in all uneven waves including Wave 7).

2.5.3.4. Wave 4

Most modules for Wave 4 have appeared in earlier waves. They include major modules on work conditions (covering, e.g., transport behaviour and job satisfaction) and modules on environmental behaviours and voluntary work. Wave 4 carries for the first time rotating modules on wealth and assets, financial attitudes and behaviours, and credit and debt. Another highlight of the Wave 4 questionnaire is a one-off module on leisure participation focussing on the Olympics 2012, which were held in and around London 27th July - 12th August 2012. The Wave 4 self-completion instrument for adults includes a focus on mental health and well-being and gender role attitudes.

The youth questionnaire instrument is the same as the one used in Wave 2 (as will be the case in all even waves including Wave 8).

2.5.3.5. Wave 5

The main difference for Wave 5 was the introduction of a set of additional consents to ask for permission to link administrative records to survey responses. These covered the domains of Higher Education and the Her Majesty's Revenue and Customs (HMRC). In addition, there was a new module around cultural participation that was asked of participants from the EMBS, GPC sample, LDA sample members and recent immigrants. An Olympics 2012 module was carried for those interviewed between the start of Wave 5 and the 26th July 2012.

The adult self-completion included new sets of questions asking about delayed self-gratification, identity and self-efficacy. There was also a set of questions for those participants in Scotland, which asked about the Scottish Referendum. Those adults issued to the second year of Wave 5 (2014) and interviewed before 22nd August were asked whether they would be voting in the referendum, and if so how they were planning to vote. Those who were interviewed after 18th September we asked whether they had voted in the referendum, and if so, how they had voted. In addition, participants were asked about mode preference; asking how likely they would be to participate in an online survey.

2.5.3.6. Wave 6

The Wave 6 household questionnaire included the additional questions on material deprivation, child deprivation, and pensioner deprivation that were previously carried at Wave 4, as well as the questions on neighbourhood conditions that were last carried at Wave 3.

The rotating content in the adult questionnaire at Wave 6 includes modules on commuting behaviour (last asked at Wave 4), work conditions (Wave 4), charitable giving (Wave 4), personal pensions (Wave 4), savings (Wave 4), Britishness (last asked of the whole sample at Wave 1), local neighbourhood belonging/service use/quality (Wave 3), membership of groups and organisations (Wave 3), political engagement (Wave 3), political efficacy (Wave 3), news and media use (Wave 3), social networks (Wave 3), three best friends (Wave 3), voluntary work (Wave 4), domestic division of labour (Wave 4), and transport behaviour (Wave 4).

In Wave 6, there were separate questionnaires for the IEMBS. The content mirrored very closely the questionnaires for the rest of the sample but due to budget constraints some modules were excluded including the youth questionnaire. IEMBS-specific new content includes the last job worked in the country of birth and first job worked in the UK after arrival, including for the respondent's parents. For a complete list see the Ethnicity and Immigration User Guide (McFall, Nandi et al. 2018).

2.5.3.7. Wave 7

Most modules in Wave 7 appeared in previous waves there are, however, a number of new modules, including questions on health service use, young adult and parental higher education expectations, and questions about poverty and shame.

Rotating content in the adult questionnaire includes the family networks, parents and children, partner relationships, child maintenance, and part of the financial behaviour and attitudes modules (all last carried in Wave 5). In Scotland, the Olympic games/commonwealth games module was included (also last carried in Wave 5). The General election module was included for those respondents who were interviewed after the election in 2015. Note that the Wave 7 questionnaire did not include questions on the referendum on leaving the European Union as there was not enough time to change the questionnaire.

The IEMB sample members now receive the same questionnaire as the main sample, but with the "Extra 5 minutes" of questions, which are also answered by the GPC and EMB samples and all non-UK born sample members in the GPS.

2.5.3.8. Wave 8

Wave 8 includes a set of modules that are carried every two waves, and so were last carried at Wave 6, such as: commuting behaviour; work conditions; charitable giving; personal pensions; savings; voluntary work; domestic labour; and transport behaviour. It also carried an identity module which is asked every three waves, so last asked at Wave 5, and a set of modules about wealth, assets and debt which are carried every four years, previously carried at Wave 4. The module on poverty shame was asked at Wave 8 for the second wave in a row. The rotation in of the identity module for the whole sample coincided with a longer identity module which was just asked of the ethnic minority boost and GPC samples. Two requests for consent to link administrative data to survey responses were made; to data held by the HMRC and data about energy use held by Department for Energy and Climate Change (DECC).

2.5.4. CHANGES TO THE QUESTIONNAIRE

All survey instruments are carefully tested in a pilot so that any issues with question wording and routing, interview flow and timings can be identified and fixed before the survey is rolled out to the main sample, see Section 2.3.2.2. The survey instruments for a wave of data collection are then fixed for the entire fieldwork period and any question (including its routing) is repeated each wave of data collection, as applicable. However, under certain circumstances changes to the questionnaire may be undertaken.

2.5.4.1. Within a Wave

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, e.g., cutting of the employment history module. At the same time other modifications were made, e.g., in question format. Notes about these changes have been documented in the variable view of the online documentation system. Also see Section 3.2, below.

2.5.4.2. Across Waves

There sometimes are changes to the questionnaire across waves. This is the case, for instance, when routing errors only became known after data collection had 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 onward this information is available for adult and proxy respondents). Another example is the SF 12 which was asked in the main interview with adults in Wave 1 but was shifted to the adult self-completion in Wave 2. Notes about these changes are documented in the variable view of the online documentation system. Also see Section 3.2, below.

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

2.6. OTHER FIELDWORK MATERIALS

Other fieldwork materials are also on the website: https://www.understandingsociety.ac.uk/documentation/mainstage/fieldwork-documents.

One example is Show cards, which are used to help respondents with their answers. Show cards are referenced in the questionnaire. Project Instructions were prepared for interviewer training and to serve as a resource in data collection. Documents for communicating with participants are also included on this portion of the website. In Wave 1, we asked for consent to link to administrative health and education records. The information leaflets and consent forms are in this section of the Study website.

The Address Record Form (ARF) is an important document for recording information about responding and non-responding households, used by NatCen in Waves 1 to 5. There are many different versions in Wave 1. Interviewers record the call record, observations on characteristics of accommodation and households, and household outcomes on the ARF. In Wave 1 there were several different versions of the ARF. The first distinction is between the GPS and the EMBS. The versions labelled ARF EB, for the EMBS, are longer because they include questions for screening household members for eligibility. ARF's labelled 2 or 3 are for addresses with multiple households and/or dwelling units. Finally, there are versions for ARF EB1 Year 1 or Year 2. This change in form was required by the change in selection criteria implemented in Year 2 of Wave 1, see Berthoud, Fumagalli et al. (2009) for more detail. The ARF screening card was a show card used during the screening interviews. Additional information about completion of the ARF can be found in the Project Instructions for Interviewers.

From Wave 6 onward, with TNS BMRB (now Kantar Public) responsible for fieldwork, the paper ARF was no longer used, and replaced by an electronic sample management system. The information continues to be stored in the ***w*_hhsamp** data files. Note that information from the ARF is not available for households that were issued for web interviews (see ***w*_issue_mode** on ***w*_hhsamp** from Wave 7 onward).

3. UNDERSTANDING SOCIETY DATA

3.1. INFORMATION ABOUT DATA FILES

The data release consists of multiple files in SPSS or Stata formats distributed by the UK Data Service. Data for different waves are released in separate files. File names begin with a prefix designating the wave of data collection ("a_" for the first wave, "b_" for the second wave; in this user guide we have used "*w*_" to denote waves in general). Data collected from different sources (i.e., e.g., the household interview, the adult interview, the youth interview) are stored in separate files. The root filename is fixed over time. For example, individual level data collected from interviews with responding adults in Wave 1 (both years: 2009 and 2010) is stored in the file **a_indresp** and individual level data collected from interviews with responding adults in Wave 2 (both years: 2010 and 2011) is stored in the file **b_indresp**.

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, the file stem name will match. Wave-specific harmonised BHPS files can be identified by the wave prefix *bw_*; files that only exist in the BHPS (or that have not yet been considered for harmonisation) have the suffix *_bh*. We refer users to the designated *Understanding Society* harmonised BHPS User Guide for more detailed information about these files, see Fumagalli, Knies et al. (2017).

Table 27 lists the data files that contain substantive information collected in interviews with responding households and individuals. They are the most likely data files analysts will want to access.

Table 28 and Table 29 list some additional data files analysts may need to access. In particular, we would like to point users to the data file **xwavedat**, which contains stable characteristics of individuals, such as ethnicity, which is typically collected only once in the lifetime of the Study. This file now includes all sample members ever enumerated in either of the studies and variables have been harmonised across studies where possible.

Table 27: List of select data files: Data from responding sample members

Filename	Description
*w*_hhresp *bw*_hhresp	Substantive data from responding households
*w*_indresp *bw*_indresp	Substantive data for responding adults (16+) including proxies and telephone interviews from individual questionnaires including self-completion
*w*_youth *bw*_youth	Substantive data from youth questionnaire (UKHLS: age 10-15, all waves; harmonised BHPS: age 11-15, Waves 4-18 only)

Table 28: List of select data files: Data from enumerated sample members

Filename	Description
*w*_indall *bw*_indall	Household grid data for all persons in household, including children and non-respondents
*w*_child	Childcare, consents and school information of all children in the household
*w*_egoalt *bw*_egoalt	Kin and other relationships between pairs of individuals in the household

Table 29: List of select data files: Cross-wave files

Filename	Description
xwavedat	Stable characteristics of individuals
xivdata xivdata_*bh*	Interviewer characteristics
xwaveid xwaveid_*bh*	Individual and household identifiers across all waves

The complete list of files and their descriptors can be seen in the online documentation system: https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation.

3.1.1. PARADATA

Some paradata, i.e., additional data collected about the interview process is available, and Table 30 lists the files in which the data are stored. Paradata consist of call records, timings data and other information collected by the interviewers during the face-to-face (or telephone) interview. The ***w*_callrec** data file has information on the number of calls made as well as the issue number, time and date and the outcome of each call. Information on the date of receipt of the case and the interviewer associated with each issue as well as the outcome at the end of each issue period is available in the file ***w*_issue**. This information and these files do not exist for the harmonised BHPS (which started out as a PAPI interview). In addition to this, information collected in the address response form (ARF) by interviewers while contacting each household and asking household members to participate in the survey is available in ***w*_hhsamp** (harmonised BHPS: ***bw*_hhsamp_bh**). This includes data on the area surrounding the address, the type of accommodation and other information that the interviewer can observe about sampled addresses.

Table 30: List of select data files: Paradata

Filename	Description
*w*_hhsamp *bw*_hhsamp_bh	Data from Address Record File for issued households
*w*_callrec	Information about interview outcome at each call
*w*_issue	Information about interview outcomes at each issue including interviewer number

Reasons for refusal are also available. Interviewers also collect some information about the quality of the interview and persons present during the interview process. This is available along with substantive data collected during adult individual interviews (including proxy interviews) in ***w*_indresp** (harmonised BHPS: ***bw*_indresp**).

3.2. INFORMATION ABOUT VARIABLES

This section relates to variables in the main Understanding Society files but many of the general principles also apply to variables in the harmonised BHPS files. A key difference is that analysists can draw upon both the BHPS documentation and the UKHLS documentation for information about variables. The UKHLS online data documentation includes the PDF versions of the questionnaires used in the BHPS Waves 1-18 but there is no hmtl version. Please see Fumagalli, Knies et al. (2017) for information about variables in the harmonised BHPS.

3.2.1. LEARNING ABOUT THE STUDY VARIABLES

There are multiple resources for learning about the Study variables in order to plan analyses. These include the questionnaires and the module and variable views in the online documentation system.

Many of the basic (non-derived) variables can be learned about directly from the questionnaires. As was shown in Figure 2, the questionnaire has much useful information. Please note that in the questionnaire, the variable name does not have the wave prefix. It also shows the brief variable label, text of the question, source of the question and value labels. Show-cards to help the respondent in answering are also marked as part of the questionnaire. You can go back and forth from the question view to the variable view.

3.2.2. VARIABLE NAMING AND LABELLING CONVENTIONS

Most variables have a mnemonic name. Variables begin with a prefix designating the wave of data collection ("a_" for the first wave, "b_" for the second wave; in this user guide we have used "*w_*" to denote waves in general. For the harmonised BHPS the generic wave prefix is "*bw_"*). 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" (cf. **Error! Reference source not found.** for lists of variables without these prefixes). 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**.

Most added-value variables, i.e., 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". The prefix "ff_" following the wave prefix shows variables that were fed forward from previous waves to route respondents appropriately in the script.

When designing *Understanding Society* we have attempted to keep the names of variables that came from the BHPS the same, for the convenience of analysts, but this has not always been possible given overruling naming conventions in the UKHLS. In producing the harmonised BHPS, the UKHLS naming conventions have been more stringently applied, see Fumagalli, Knies et al. (2017).

Note that the variable name does not change over time so long as the underlying question does not change substantially. Analysts are advised, however, 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.

3.2.3. VARIABLE VALUES AND LABELS

The detailed variable view provides information about valid and invalid responses. Additional codes denote different types of reasons for the lack of a valid response. These values have not been specified as missing in Stata or SPSS. However, these statistical packages have commands to assign values to missing for many variables simultaneously. describes the missing value codes. The meaning of other values is explained with the variable's value labels. There may also be notes in the detailed variable view of the online documentation system on the website: https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation.

Table 31: Missing value codes

Value	Description
-21	No data from the UKHLS
-20	No data from the BHPS W1-18
-11	Only available for the IEMBS
-10	Not available for the IEMBS
-9	Missing by error or implausible
-8	Not applicable to the person or because of Routing
-7	Proxy respondent. The question was not asked of proxy respondents or derived variable cannot be computed for proxy respondents.
-2	Refused
-1	Don't know

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 processing was therefore not as closely monitored, e.g., respondents may not have followed the question routing. We recommend that users carefully read the questionnaires and compare missing value distributions across waves before using the substantive information contained in them.

3.2.4. IDENTIFIERS AND POINTERS TO OTHER HOUSEHOLD MEMBERS

Households are identified 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.

Individuals are identified by the personal identifier (**pidp**), which is consistent 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 and included in the harmonised BHPS files (in addition to the former BHPS household identifier ***bw***_**hid**).

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" in the Online Dataset Documentation. See Fumagalli, Knies et al. (2017) for a description of pointers to significant others in the harmonised BHPS files.

3.2.5. STABLE INFORMATION ABOUT SAMPLE MEMBERS

The **xwavedat** file contains stable information about all sample members, which include, from the Wave 7 data release onward, all former BHPS sample members, i.e., irrespective of whether they ever participated in *Understanding Society* interviews. The variable **xwdat_dv** reports whether a sample members was enumerated only in the BHPS or UKHLS, or in both studies.

The variable listing of the **xwavedat** data file in the online dataset documentation provides a complete list of all variables included in this file. They include information collected when first entering the Study (initial conditions module) such as the respondent's first job after leaving school (**j1soc00_cc**) and information about the respondent's parents' when the respondent was aged 14 (e.g., **masoc90_cc**) which is also stored in the wave-specific **indresp** files in which the respondent answered these questions.

3.2.6. 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, whilst others are computed post-field for the purpose of analysis. The suite of derived variables included in *Understanding Society* includes flags for whether or not a certain characteristic is true for a study member (e.g., ***w*_jbft_dv** is a flag for whether or not a respondent has a full-time job), counts of the number of people in the household for whom a certain characteristic is true (e.g., ***w*_nemp_dv** is the number of employed people in the household), and pointers to significant others in the household (e.g., ***w*_mnpid** records the cross-wave person identifier of the respondent's biological mother). As a rule of thumb, variables that are derived post-field end on the suffix "_dv", and pointers to others in the household end on "pno" or "pid"; they can, therefore, easily be identified in the data. Note that all variables ending on "pid" contain the UKHLS person identifier **pidp** (i.e., not the original BHPS person identifier pid). The variable labels clearly state this. Derived variables that are created post-field are added last to the data files and can therefore also be easily identified by their position in the files; derived variables computed during the interview appear in the context of the relevant module.

Note that a data file may offer alternative versions of a derived variable. This is particularly true for derived variables that point to others in the household: One set of variables (e.g. variable names starting on "hg") has been computed based on information collected in the household grid module of the questionnaire during the interview. The alternative version is computed post-field after the information collected in the household grid has undergone extensive data cleaning. See and compare, for example, ***w*_hgbiom** and ***w*_mnpno** for the person number of the respondent's biological mother in the household.

From Wave 2 onward, proactive dependent interviewing was used to increase efficiency of data collection and lessen respondent burden. Specifically, information reported at an earlier time is fed forward to the respondent to personalize the

question. So rather than ask a question about current occupation with its complex probing by interviewers, the question might say, “the last time you were interviewed you said you were “specific occupation” are you still “specific occupation”? 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). The variable **b_ff_plbornc** is the country of birth of the respondent fed forward from the previous wave. Note the use of the prefix “ff_”. Some of the fed-forward variables were not used in the wording of a question but were used by the CAPI script to route respondents appropriately based on information from the previous wave.

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 (i.e., the variable stem name from Wave 1). See, for example, the socio-economic classification of the current job (***w*_jbsoc00**) and the standard industrial classification (***w*_jbsic07**).

We use look-up files between SOC 2000 and other classifications to derive additional occupational classifications. Users may apply to access the Special Licence version of *Understanding Society* to access non-condensed versions of these codes which will allow them to derive additional classifications (such as EGP using the Stata command –isko-).

Information about how the derived variable is produced is shown in the notes for derived variables in the detailed variable view of the online documentation. The view provides descriptive statistics and, in the Origin field, lists of 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 should consult the description of any (derived) variables that they plan to use in their analysis. Derived variables are listed under the Index Term “Derived variables” in the Online Dataset Documentation.

3.2.7. SAMPLE DESIGN VARIABLES

As the sample design involves stratification, clustering and weighting, 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. The weighting variables are described in Section 3.3. Here we describe the stratification and clustering variables in the main UKHLS files. See Fumagalli, Knies et al. (2017) for a description of these variables in the harmonised BHPS. More detailed information about survey design and weights in the (harmonised) BHPS can be found in Section A5 1-13 in Taylor (2010). General advice on using this information appropriately applies to both the UKHLS and harmonised BHPS data.

The variable indicating the primary sampling unit is ***w*_psu.** This is an indicator of the primary sampling unit (PSU) to which the sample member belongs. The prefix “*w*_” denotes waves in general. The value of ***w*_psu** does not change between waves, but for new sample entrants it is only defined from the wave at which they enter the sample. Values on the variable ***w*_psu** are further described in Table 32.

The variable indicating the strata is called ***w*_strata**. This indicates the sampling stratum from which the sample member was selected. The value of ***w*_strata** does not change between waves, but for new sample entrants it is only defined from the wave at which they enter the sample. Table 33 lists the range of values on ***w*_strata**.

Table 32: Description of *Understanding Society* Primary Sampling Unit variable

Value	Sample	Notes
1 – 575	former BHPS sample in GB	Identical to the BHPS variable ***w*psu**
701 – 1999	former BHPS Northern Ireland sample	Corresponds to initial (BHPS Wave 11) sampled households, as these were selected in a one-stage design
2001 – 4640	UKHLS-GPS in England, Scotland and Wales	Corresponds to the postal sectors used as PSUs, see Lynn (2009)
46424 – 7035	UKHLS-GPS in Northern Ireland	Corresponds to Wave 1 sampled households, as these were selected in a one-stage design
7048 – 51789	UKHLS-EMB	Corresponds to Wave 1 sampled households, as these were selected in a one-stage design within the high minority density domain, see Berthoud, Fumagalli et al. (2009)
52001 – 52250	UKHLS – IEMBS	Corresponds to the 250 small areas selected as PSUs (LSOAs in England and Wales, Data zones in Scotland)

Note: there was an error in **b_psu** and **c_psu** for Northern Ireland BHPS households in the Wave 2 and Wave 3 data releases. This has been corrected from the Wave 4 release.

Table 33: Description of *Understanding Society* stratification variable

Values	Sample	Notes
1 – 151	former-BHPS sample in GB	Identical to the BHPS variable ***w*strata**
701	former-BHPS Northern Ireland sample	Northern Ireland treated as a single stratum
2001 – 3320	UKHLS-GPS in England, Scotland and Wales	Corresponds to groups of two or more PSUs in selection order, as they were selected systematically from an implicitly ordered list, see Lynn (2009)
3321	UKHLS-GPS in Northern Ireland	Northern Ireland treated as a single stratum
3322 – 5117	UKHLS-EMB	Corresponds to the postal sectors in the high minority density domain as selections were made independently from each, see Berthoud, Fumagalli et al. (2009)
5121 – 5124	UKHLS-IEMBS	Corresponds to the four sampling strata

3.3. WEIGHTING ADJUSTMENTS

The dataset is designed to be used with weights. If you are planning to not use weights please state the assumptions you are making explicitly when describing results (for the assumptions please see Section 4.1).

A number of weights are provided for data users. They adjust for unequal selection probabilities, differential nonresponse, and potential sampling error. 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.

Separate sets of weights are provided for

- the former-BHPS sample,
- the combined GPS and EMBS components,
- the combined GPS, EMBS and BHPS components
- the combined GPS, EMBS, BHPS and IEMBS components (from Wave 6).

The available sets of weights are not identical for these three analysis bases, reflecting differences in data collection. Considering the complexity of the Study design, weights should be selected carefully, following the advice provided below.

The first part of this section covers the purpose of the weights and how to use the naming conventions for the weight variables to interpret and select the different weight variables from among a complex assortment. **It is the most important section for a user wanting to select the appropriate weight for a planned analysis.** Section 3.3.3 then provides the technical details of how weights were calculated.

If your aim is to generalise to the UK population, unweighted analyses should be avoided. For advanced users who want to model nonresponse in their own way, we provide design weights and inclusion weights (see below) which adjust the sample for unequal selection probabilities. 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.

Note that a number of longitudinal weights are provided corresponding to the year of sample selection (BHPS since 1991; BHPS since 2001; GPS and EMBS since 2010-2011; combined BHPS, GPS and EMBS since 2011-2012; and combined BHPS, GPS, EMBS and IEMBS since 2014-2015). Cross-sectional weights from Wave are based on combined BHPS, GPS and EMB samples. From Wave 6, additional cross-sectional weights are based on combined BHPS, GPS, EMBS and IEMBS.

Here we describe the weights in the main UKHLS files. Full details on survey design and weights in the (harmonised) BHPS can be found in in Taylor (2010), Section A5 1-13. Some information, in a nutshell, is provided also in Fumagalli, Knies et al. (2017). The general advice we provide on using the survey design variables and weights appropriately applies to both the UKHLS and harmonised BHPS data.

3.3.1. SELECTING THE CORRECT WEIGHT FOR YOUR ANALYSIS

Given the complexity and multi-purpose nature of the *Understanding Society* design we provide multiple weights to meet the different needs of users. The weight for your

analysis reflects the survey instrument, which is the source of the data being used in the analysis, the analysis level (household or individual), and the combination of waves involved.

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

The naming conventions for weights are intended to help users to pick 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). The rules are described in the "Naming Conventions for Weighting Variables" section below.

If your analysis uses only data from Wave 4, select the "xw" (cross-sectional) version of the weight. This weight is defined for all sample members who responded to the relevant survey instrument at Wave 4. If your analysis uses data from multiple waves select an appropriate "lw" (longitudinal) version of the weight.

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

For example, if in one cross-sectional model for Wave *n* you use questions from the proxy and full interview as well as from the self-completion, then the correct weight will be ***n*_indscus_xw** – the weight for the self-completion questionnaire, as its level (1) is lower than the level for proxy and full interview (3).

Table 34: Selecting the correct weight: Hierarchy of analysis levels

Level of Analysis	Questions available for
4	household level (all enumerated individuals)
3	Adult proxy and main interview
2	Adult main interview only (no proxy)
1	Adult or youth self-completion interview

The following tables list the weight variables. The list has been broken into separate tables so the user can go quickly to the data source for the planned analysis and then select the particular relevant weight, for example cross-sectional vs. longitudinal. Each table focuses on a major data source and has the weight variables used for cross-sectional and longitudinal analyses. Please also note that weights are defined for particular sample components. Wave prefixes refer to a specific wave (a_ or b_) or to waves in general *n_.*

Here is an example: a longitudinal weight for BHPS + EMBS + GPS represents Waves 2 and 3, but the first time one can find it is in the Wave 3 release. Thus, waves representing is 2-3, but the wave it starts from is 3. We would indicate that the above weight represents waves 2+ (all waves starting from 2).

Where is this useful? If a user wants to analyse longitudinal data starting with Wave 1, then they cannot use the combined longitudinal weight (for BHPS + GPS + EMBS), but will have to use the weight for GPS +EMBS (by looking at 'wave

representing' they should be able to find quickly which weight is best for their analysis).

Note that the weights listed in Table 35 through Table 38 all apply at the individual level.

Table 35: Weight variables for analyses using household grid or household interview

Analysis level	Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
Household	1		Household grid and/or household interview	a_hhdenus_xw
Household	*N*	2	Household grid and/or household interview (BHPS, GPS and EMB)	*n*_hhdenub_xw
Individual	1		Household grid and/or household interview	a_psnenus_xw
Individual	*N*	2-5	Household grid and/or household interview (BHPS, GPS and EMB)	*n*_psnenub_xw
Individual	*N*	6	Household grid and/or household interview (BHPS, GPS, EMBS and IEMBS)	*n*_psnenui_xw
Individual	1+	2	Household grid and/or household interview (GPS and EMB)	*n*_psnenus_lw
Individual	1991+	2	Household grid and/or household interview (BHPS, GB 1991)	*n*_psnen91_lw
Individual	2001+	2	Household grid and/or household interview (BHPS, UK 2001)	*n*_psnen01_lw
Individual	2+	3	Household grid and/or household interview (BHPS, GPS and EMBS since 2010-2011)	*n*_psnenub_lw
Individual	6+	7	Household grid and/or household interview (BHPS, GPS, EMBS and IEMB since 2014-2015)	*n*_psnenui_lw

Note: Data users should be aware of a change in the definition of the cross-sectional population represented by the Study at Wave 6. From Wave 6, the sample is representative of people who have lived continuously in the UK since 2014-15, i.e., recent immigrants are now included. We suggest that all cross-sectional analysis uses the weight *n*_psnenui_lw which incorporates these new immigrants. At Waves 1 to 5, the sample is representative of people who have lived continuously in the UK since 2009-10. For special situations when a user wants to represent this population cross-sectionally in later waves the weight *n*_psnenub_xw should be used.

Table 36: Weights for analysis using adult main and proxy interviews

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult main and proxy interview	a_indpxus_xw
2		Adult main and proxy interview	b_indpxus_xw
2		Adult main and proxy interview (BHPS)	b_indpxbh_xw
N	2 - 5	Adult main and proxy interview (BHPS, GPS and EMBS)	*n*_indpxub_xw
N	6	Adult main and proxy interview (BHPS, GPS, EMBS and IEMBS)	*n*_indpxui_xw
1+	2	Adult main and proxy interview	*n*_indpxus_lw
2+	3	Adult main and proxy interview (BHPS, GPS and EMBS since 2010-2011)	*n*_indpxub_lw
6+	7	Adult main and proxy interview (BHPS, GPS, EMBS and IEMBS since 2014-2015)	*n*_indpxui_lw

Table 37: Weights for analysis using adult main interviews

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult main interview	a_indinus_xw
2		Adult main interview	b_indinus_xw
2		Adult main interview (BHPS)	b_indinbh_xw
n	2 – 5	Adult main interview (BHPS, GPS and EMBS)	*n*_indinub_xw
n	6	Adult main interview (BHPS, GPS, EMBS and IEMBS)	*n*_indinui_xw
1+	2	Adult main interview	*n*_indinus_lw
1991+	2	Adult main interview (BHPS, GB)	*n*_indin91_lw
2001+	2	Adult main interview (BHPS, UK)	*n*_indin01_lw
2+	3	Adult main interview (BHPS, GPS and EMBS since 2010-2011)	*n*_indinub_lw
6+	7	Adult main interview (BHPS, GPS, EMBS and IEMBS since 2014-2015)	*n*_indinui_lw

Table 38: Weights for analysis using adult "Extra 5 minutes" interview

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult "Extra 5 minutes" interview	a_ind5mus_xw
n	3	Adult "Extra 5 minutes" interview	*n*_ind5mus_xw
1+	2	Adult "Extra 5 minutes" interview	*n*_ind5mus_lw

Table 39: Weights for analysis using adult self-completion

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Adult self-completion	a_indscus_xw
2		Adult self-completion	b_indscus_xw
2		Adult self-completion (BHPS)	b_indscbh_xw
n	2 - 5	Adult self-completion (BHPS, GPS and EMBS)	*n*_indscub_xw
1+	2	Adult self-completion	*n*_indscus_lw
2+	3	Adult self-completion (BHPS, GPS and EMBS)	*n*_indscub_lw
6+	7	Adult self-completion (BHPS, GPS, EMBS and IEMBS)	*n*_indscui_lw
Note: A cross-sectional weight for the self-completion component was discontinued after Wave 5 as levels of nonresponse to this component are low. This can be treated in the same way as item nonresponse and the relevant individual interview cross-sectional weight can be used when analysing data from the self-completion component.			

Table 40: Weights for analysis using youth self-completion

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
1		Youth self-completion	a_ythscus_xw
2		Youth self-completion	b_ythscus_xw
2		Youth self-completion (BHPS)	b_ythscbh_xw
n	3	Youth self-completion (BHPS, GPS and EMB)	*n*_ythscub_xw
n	7	Youth self-completion (BHPS, GPS, EMB and IEMBS)	*n*_ythscui_xw

Table 41: Weights for analysis using nurse health assessment data

Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight
2+	3	Nurse visit with Wave 2 to Wave *n* full interviews (GPS, BHPS)	*n*_indnsub_lw
2+	3	Nurse visit with Wave 2 to Wave *n* full interviews and consent to blood sample at Wave 2 (GPS, BHPS)	*n*_indbdub_lw
1991+	3	Nurse visit with 1991 to Wave *n* full interviews (BHPS)	*n*_indns91_lw
1991+	3	Nurse visit with 1991 to Wave *n* full interviews and consent to blood sample at Wave 2 (BHPS)	*n*_indbd91_lw
Note: Other weights for use with the health assessment data that were collected at Waves 2 and 3 are available with the separate health assessment data release, see Section 5.3.2, and are fully documented in the associated user guide, see McFall, Petersen et al. (2014). These four weights are documented here as they will be released with the main data release each year.			

Table 42: Design and inclusion weights

Analysis level	Wave(s) / Years representing	Wave starts from	Data source	Analysis Weight*
Household			(Wave 1 household design weight)	a_hhdenus_xd
Household			(Wave 1 household design weight for GPS only)	a_hhdengp_xd
Individual			(Design weight)	a_psnenus_xd
Individual			(Design weight GPS only)	a_psnengp_xd
Individual			("Extra 5 minutes" design weight)	a_ind5mus_xd
Individual			BHPS inclusion weight for OSMs issued into UKHLS	b_psnenbh_li
Individual			BHPS-2010 longitudinal enumerated person weight	b_psnenbh_lw
Individual			BHPS, GPS and EMBS combination inclusion weight	b_psnenub_li
Individual			BHPS, GPS, EMBS and IEMBS combination inclusion weight	f_psnenui_li
Notes: *Typically for advanced users (see technical details).				

3.3.2. NAMING CONVENTIONS FOR WEIGHTING VARIABLES

The naming conventions will help users to select the weight they need or to interpret the purpose of some weight variables. The structure of variable names for weights takes the format wave prefix followed by target population, survey instrument and weight type, or *w*_xxxyyzz_aa.

Table 43 presents the available options for these components.

Table 43: Naming convention for *Understanding Society* weights

*w*_	Xxx	Yy	Zz	_aa
a_ **b**_ **c**_ **d**_ …	**hhd**: household **psn**: persons 0+ **ind**: persons 16+ **yth**: persons 10-15	**en**: enumeration **in**: interview **px**: interview or proxy **5m**: "extra 5 minutes" **sc**: self-completion **ns**: nurse visit **bd**: blood	**us**: GPS & EMB **bh**: BHPS **ub**: GPS, EMB & BHPS **ui**: GPS, EMB, BHPS & IEMB **91**: BHPS original sample **01**: BHPS original sample + boosts	_**xw**: cross-sectional analysis weight _**lw**: longitudinal weight _**xd**: x-sectional design weight _**li**: longitudinal inclusion weight
* "gp" letters are used for weights available for the GP sample only. But there is only type of such weight - the design weights for the GP sample. This weight should be used by advanced users only. For all cases we advise you not to use the GPS sample by itself.				

For example, **a_indinus_xw** is the cross-sectional analysis weight for individual interview data from Wave 1, representing the population of persons aged 16 or older.

b_indscus_lw is the longitudinal analysis weight for individual self-completion interviews from Wave 1 and Wave 2 representing the adult population who continuously lived in UK at the time of Wave 1 and 2.

3.3.3. TECHNICAL DETAILS

In this section we describe in turn how the weights were derived for:

- GPS and EMBS Wave 1 weight;
- GPS and EMBS longitudinal weights;
- BHPS longitudinal weights;
- Combined sample (BHPS, GPS and EMBS) longitudinal weights;
- Combined sample (BHPS, GPS and EMBS) cross-sectional weights.

3.3.3.1. *Wave 1* (GPS and EMBS) Weights

The Wave 1 household level weights consist of two components: a design weight and nonresponse adjustment for household level nonresponse. Wave 1 individual

level weights consist of four components: the design weight, nonresponse adjustment for household level nonresponse, individual level within-household nonresponse, and post-stratification to population characteristics. Each of the components is explained below.

Design Weight

The design weight corrects for unequal probability of selection at a number of levels.

The household level design weight corrects for:

- Unequal selection probability due to the boost in Northern Ireland. The GPS selection probabilities in Northern Ireland are approximately twice those in other parts of the UK;
- Unequal selection probability related to selection into the EMB. Selection probabilities in the EMBS part of the sample vary considerably between areas, depending on the estimated ethnic mix of the area and ethnic composition of the household. Additionally, households in high density areas with at least one ethnic minority member were weighted to account for combined probability of being selected as part of the GPS or as part of the EMBSs;
- The selection probability of households in a dwelling with more than three households or at an address with more than three dwellings is adjusted for the fact that only three such households were selected from the same address.

Individual level design weights correct for all the above with one specific difference: non- ethnic minority persons who live with ethnic minority persons in the same household have a chance to be selected only via the GPS part of the sample, and not via the EMBS. This means that non- ethnic minority persons in the EMBS (who are TSMs) are given a design weight of 0 while non- ethnic minority persons in the GPS are given the household design weight. The weights for ethnic minority persons are adjusted for their dual probability of being part of GPS and EMBS.

Individual level design weights for those eligible to answer the “Extra 5 minutes” is similar to the above design weight but differs in the following ways. It adjusts for the fact that the GPC sample is only 1/45th of the GPS original sample; that all ethnic minority members in low-density areas were administered the “Extra 5 minutes”; and that ethnic minority members in high-density areas had a chance to be selected into either the GPC sample or the EMBS. Similar to the above weight, non- ethnic minority persons were assumed to have a chance to be part of the GPC sample only and not part of the EMBS.

Additionally, we provide GPS design weights (**a_hhdengp_dw** and **a_indengp_dw**). These weights are valid only for sample members selected through the GPS, and adjust for oversampling in Northern Ireland and for subsampling within households from multiple dwellings per address or multiple households per dwelling.

Household-level Nonresponse Adjustment

Household level nonresponse adjustment is more complex than in other surveys given the large number of households which were selected as part of the EMBS with unknown eligibility. Households who were selected as part of the EMBS were screened on whether they contain at least one member of a relevant ethnic minority

group (Berthoud, Fumagalli et al. 2009). Given the low proportion of eligible households in the EMBS it is unrealistic to assume that all non-responding households would be eligible, that is, contain at least one ethnic minority member. To take this into account we modelled eligibility and used this information in household nonresponse adjustments such that households which were more likely to be eligible had a higher influence on the nonresponse correction. Note, that the predicted eligibility multiplied by the design weight is released for all the EMBS households of unknown eligibility as part of **a_hhdenus_xd**. This will enable an advanced user to model Wave 1 household nonresponse taking into account the chance to be eligible among households of unknown eligibility.

To model eligibility we used predictors from the sampling frame and administrative neighbourhood data linked at a geographical level (for detailed description see below). After excluding ineligible addresses (like businesses or demolished and non-existent addresses), the eligibility was modelled using only EMBS households with known eligibility status (either screened out or screened in). This prediction was then extrapolated onto EMBS households of unknown eligibility (not contacted). Given the limited number of selected addresses in Wales and Scotland and differences between countries in the available auxiliary variables (see below), we predicted eligibility using two models. The first included common predictors for England and Wales and eligibility was predicted for these two countries. The second was based on England, Wales and Scotland, using a more limited number of predictors. Eligibility for Scotland was predicted only from this model.

Following this, the probability of responding was estimated using backward stepwise logistic regression weighted by eligibility status (where the ineligible were excluded, those known to be eligible had an eligibility of one, and those with unknown eligibility had a weight proportional to the predicted probability of being eligible obtained from the above model). The predictors used in this model were the same as for the eligibility model and are described in detail below. Given that administrative neighbourhood data differs between England, Wales, Scotland and Northern Ireland, a separate model was implemented for each country. GPS and EMBS response propensity was modelled together (which allowed us to model nonresponse within each country separately), but the indicator of EMBS was retained in the model even if it was not statistically significant.

Predictors used for eligibility model and household level nonresponse correction come from the following sources:

- Sampling frame information, including such variables as sample month and geographical region;
- Predicted ethnic density of the postcode sector for five main ethnic groups in England, Scotland and Wales, as described in Berthoud, Fumagalli et al. (2009);
- A wide range of indicators from Census 2001 and the most updated version of neighbourhood statistics as of summer 2011, linked separately for England, Wales, Scotland and Northern Ireland (see below).

The household nonresponse correction weight was calculated as the inverse of probability from the above model. This weight was multiplied by the household design weight to create the Wave 1 household level weight. The design effect was

estimated using this weight. No truncation was necessary. The obtained weight was scaled to a mean of 1 and was named **a_hhdenus_xw**.

Neighbourhood Statistics

For England and Wales the information was linked at Middle Layer Super Output Area (MSOA) or Lower Layer Super Output Area (LSOA) levels and was obtained from http://neighbourhood.statistics.gov.uk. The examples of linked information obtained from Census 2001 include the proportions in the MSOA of employed, retired, outright property owners, travellers to work using different types of transport, single household members, households with one car, people with different types of qualification and professional occupation, among others. Other linked information includes 2010 information on multiple deprivation indexes, on crime instances, 2009 information on inflow and net change of neighbourhood population, the proportion of different allowance claimants, and 2008 information on hospital admissions and energy consumption.

For Scotland the information was linked at the data zone level from http://www.scotlandscensus.gov.uk/ods-web/data-warehouse.html and http://www.scotland.gov.uk/Topics/Statistics/SIMD. From the Census 2001, information was obtained on population density, mean age, average household size, and number of rooms per household in the data zone, as well as the proportions in the data zone born in Scotland and outside the EU, of different religious denominations, employed, unemployed and retired, disabled, those with different levels of qualification and types of occupation, and different types of accommodation, among others.

For Northern Ireland the information was linked at the Super Output Area (SOA) level and was obtained from http://www.ninis.nisra.gov.uk. Examples of predictors obtained from Census 2001 at the SOA level include the average hours worked by residents, the average age of residents, percentages of residents with different level of qualifications, with different employment statuses, and with different types of marital status, among others. The predictors also include 2007-2009 information on multiple deprivation indexes.

Note, that using Understanding Society analysis weights (all but design weights), adjusts for household nonresponse bias in any estimate, to the extent it is related to the above mentioned variables.

Enumerated Individual Weight

The weight for analysis of enumerated individuals (**a_psnenus_xw**) is not equivalent to the household weight for all household members, as often happens in other household studies. This is because we have TSMs in Wave 1, who are not ethnic members, selected into EMBS part of the sample. Thus, the individual level design weight is not equal to the household level design weight for individuals in households containing a mix of ethnic minority and non-ethnic minority persons. The weight for the analysis of enumerated individuals is calculated as the product of the individual level design weight **a_psnenus_xd** and the household level nonresponse correction (described above). The design effect was tested showing that no truncation was necessary. Weighted sample distributions were then compared to ONS mid-year

estimates (with a correction for institutionalized population) and post-stratification was implemented for the fully crossed matrix of gender by geographical region by 5-10 year age groups. Thus the individual level enumerated weight consists of:

> The individual level design weight multiplied by the household nonresponse correction, multiplied by the post-stratification adjustment. The obtained weight is scaled to have a mean of one.

Individual-level Nonresponse Adjustment

Five different individual level weights were prepared for users reflecting nonresponse occurring at different levels and different questionnaire instruments. Each individual level weight consists of this product:

> The individual level design weight, multiplied by the household nonresponse correction, multiplied by the individual level nonresponse correction conditional on household response, multiplied by the post-stratification adjustment.

The individual nonresponse correction (conditional on household nonresponse) is modelled at three levels:

- For adult respondents (age 16 or older) who either completed the main interview or for whom a proxy interview was completed (for **a_indpxus_xw**);
- For adult respondents (age 16 or older) who completed the main interview only (for **a_indinus_xw** and **a_ind5mus_xw**);
- For respondents aged 10 or older who completed and returned the self-completion questionnaire (for **a_indscus_xw** and **a_ythscus_xw**).

Note, that the same model was used for respondents regardless of whether they were selected into GPS or EMBS; that the response propensity is assumed to not depend on whether respondents received the “Extra 5 minutes” or not; and that conditional on age (present in the model), the response to self-completion is assumed to have the same predictors for adults and youth (this assumption allowed modelling the response in each country separately, which wouldn’t otherwise be possible for youth sample).

The individual level response, conditional on household response, was modelled using backward stepwise logistic regression separately for England, Wales, Scotland and Northern Ireland. The four models were implemented for each of the three levels described above. The predictors used in the models include all the predictors used for the household level nonresponse models and individual and household-level variables obtained from the household questionnaire, such as age and gender, marital and employment status, household size and presence of children in the household, as well as household expenditure on food and food outside, consideration of use of environmental energy, among others.

The individual-level nonresponse adjustment was obtained as the inverse of the predicted probability and was then multiplied by the relevant (either individual or “Extra 5 minutes”) design weight and by the household nonresponse correction. No truncation was deemed necessary as there were no extreme values substantially impacting the design effects. The post-stratification was implemented as described above in the individual level enumeration weight section, except that a greatly reduced matrix was used in the case of the “Extra 5 minutes” weight, due to the

much smaller sample size for which this weight applies. After multiplying by the post-stratification adjustment, each of the obtained weights was then scaled to a mean of one.

3.3.3.2. GPS and EMBS Longitudinal Weights

Each of the five types of longitudinal weights (enumerated persons, proxy or main interview, main interview, self-completion, and "Extra 5 minutes" interview) is based on the corresponding previous longitudinal weight (except in Wave 2 where it is based on Wave 1 cross-sectional weight). An additional adjustment for nonresponse since the last wave is applied. Each adjustment is based on a model of Wave *n* response conditional on Wave (*n*-1) non-zero longitudinal weight for the instrument in question. For the enumerated person model, covariates are taken from the Wave (*n*-1) household grid and household questionnaire. In the model for proxy and main interviews, covariates were taken from the Wave (*n*-1) proxy interview (or the equivalent items from the main interview), household grid and household questionnaire. In both the model for main interviews and the model for adult self-completion questionnaires, covariates are taken from the Wave (*n*-1) main interview, household grid and household questionnaire. The adjustment weight is calculated as the reciprocal of the model-predicted response propensity. The Wave (*n*-1) weight is then multiplied by the Wave *n* adjustment to create the Wave *n* longitudinal weight.

Newborns born to an OSM mother since the Wave (*n*-1) interview receive the longitudinal enumerated person weight of their mother (reflecting the idea that the probability of observing the newborn is equal to the probability of observing the mother). The principle behind the longitudinal weights is that they are defined for each person who is observed at all of the relevant waves for which they were eligible. For this reason, newborns observed at Wave *n* receive a Wave *n* longitudinal weight as they were enumerated at Wave *n*, the only wave for which they were eligible.

3.3.3.3. Cross-sectional Weights (GPS and EMBS) for *Wave 2*

The cross-sectional enumerated individual weights are based on the longitudinal enumerated individual weights, which are allocated through a weight-share method to temporary sample members (TSMs) and permanent sample members (PSMs) who entered the sample at Wave 2. Note, that only new TSMs and PSMs entering the Study after Wave 1 receive a shared weight. TSMs who were present in Wave 1 (in the EMBS) are given a cross-sectional weight of 0. This is done as the GPS part of the sample does not have an equivalent TSM group (OSM non-ethnic minority members living with TSM ethnic minority members). Giving a cross-sectional weight of 0 to Wave 1 TSMs maintains the balance of the whole sample.

These cross-sectional enumerated individual weights then serve as the base for the other cross-sectional individual-level weights, each of which (main, main or proxy, self-completion, youth) involves an additional adjustment for nonresponse to the relevant instrument conditional on enumeration. The nonresponse models are therefore based on all eligible persons enumerated at Wave 2 (including TSMs and those OSMs who did not respond to the respective instrument at Wave 1), with covariates taken from responses to *Understanding Society* Wave 2 household grid and household questionnaire.

The cross-sectional weights for households (**b_hhdenus_xw**) are set equal to the minimum nonzero longitudinal enumerated person weight (**b_psnenus_lw**) amongst adults in the household, reflecting the idea that the probability of observing the household is equal to (or greater than) the probability of observing the person in the household who has the greatest probability of being observed.

3.3.3.4. BHPS Longitudinal Weights

Four weights will be continued from BHPS with changed variable names. The corresponding weight variables are:

- ***w*ewght** now called ***w*_psnen91_lw**,
- ***w*lewtuk1** now called ***w*_psnen01_lw**,
- ***w*lrght** now called ***w*_indin91_lw**, and
- ***w*lrwtuk1** now called ***w*_indin01_lw**, where ***w*** represents the most recent *Understanding Society* wave.

These weights are based on Wave 18 BHPS longitudinal weights, which account for the first wave household nonresponse, the first wave within household individual nonresponse (to enumeration or to an individual main questionnaire, respectively) and for individual nonresponse between the first wave and Wave 18 of BHPS. The base weights which reflect continuous enumeration (**rlewght**, a BHPS variable name) and continuous response to the main questionnaire (**rlrght**, a BHPS variable name) since 1991 are used for creating weights for longitudinal analysis starting 1991. Note that such an analysis excludes Northern Ireland as it was added to BHPS in 2001 and will also exclude the Scotland and Wales boost samples that were added in 1999. Similarly, the base weights which reflect continuous enumeration (**rlewtuk1**, a BHPS variable name) and continuous response to main questionnaire (**rlrwtuk1**, a BHPS variable name) since 2001 are used for creating weights for longitudinal analysis starting in 2001. Analysis using these weights will include all the BHPS samples. For more information on the BHPS weight calculation please refer to BHPS documentation (Taylor 2010).

For each of the Wave 18 weights an additional adjustment is applied to correct for attrition between Wave 18 of the BHPS and Wave 2 of *Understanding Society*, when the BHPS joined *Understanding Society.* The adjustment is the reverse of the estimated probabilities of participation (enumeration or response to main questionnaire) based on logistic regressions predicting participation at Wave 2 of *Understanding Society* conditional on participation at Wave 18 of BHPS. The covariates used in the model predicting enumeration are from the BHPS Wave 18 household grid and household questionnaire. The same covariates plus covariates from the Wave 18 main questionnaire are used for predicting response to the *Understanding Society* Wave 2 main questionnaire. Enumeration weights for newborn babies (biological, step or natural) born to an OSM mother since the time of the BHPS Wave 18 interview are equal to their mother's enumeration weight.

For "Rising 16s" (OSMs who turned 16 between the time of the BHPS Wave 18 interview and the *Understanding Society* Wave 2 interview and who could therefore be aged 16, 17, or even 18 at the time of UKHLS Wave 2), main response weights consist of the relevant longitudinal enumerated person weight, with an adjustment for the probability of main response at Wave 2 conditional on enumeration at Wave 2. The adjustment is the inverse of the response propensity predicted by a separate

logistic regression model (based just upon all adults and inferred to "Rising 16s") using covariates from the Wave 2 household questionnaire and household grid. The base weight for "Rising 16s" correction is continuous enumeration since 1991 (**b_psnen91_lw**) for the BHPS 1991 main response weight (**b_indin91_lw**), and is the BHPS-2010 longitudinal enumerated person weight (**b_psnenbh_lw** – see next section below) for the BHPS 2001 main response weight (**b_indin01_lw**). The main response weight for each rising 16 year-old is then scaled by a constant factor so that the ratio of "Rising 16s" to older adults among main questionnaire respondents equals the equivalent proportion among all enumerated respondents. The weights (**b_psnen91_lw**, **b_psnen01_lw**, **b_indin91_lw** and **b_indin01_lw**) are calculated by multiplying the respective BHPS Wave 18 weight and the adjustment, and are scaled to one.

Starting at Wave 3, longitudinal weights for BHPS are created based on the previous wave longitudinal weights. For longitudinal enumeration weights (***n*_psnen91_lw** and ***n*_psnen01_lw**), enumeration in Wave ***n*** is predicted among adults having positive longitudinal weight in previous wave. Enumeration is modelled using logistic regression with covariates from the household questionnaire and household grid from wave *n*-1. Newborns are given the enumeration weight of their mother. The enumeration weights are then scaled to a mean of one.

For longitudinal main interview weights (***n*_indin91_lw** and ***n*_indin01_lw**) response to the main interview is predicted using logistic regression with predictors obtained from the wave *n*-1 household questionnaire, household grid, and main questionnaire. The probability is then inverted. The response of "Rising 16s" is predicted in a separate logistic regression in which response to the main questionnaire for all adults 16 and over is estimated using predictors from the wave *n* household questionnaire and household grid and conditional on enumeration in current wave. The response probabilities are then inverted and multiplied by the base weights (***n-1*_indin91_lw** or ***n-1*_indin01_lw** for adults; and ***n*_psnen91_lw** or ***n*_psnen01_lw** for "Rising 16s"). The weighted ratio of "Rising 16s" to others is scaled to reflect the ratio of these age groups as estimated using the longitudinal enumeration weight in wave *n.*

3.3.3.5. BHPS Cross-sectional Weights for *Wave 2*

The BHPS cross-sectional weights are created as follows: we first model the chance of each BHPS OSM being issued into the *Understanding Society* (reflected in **b_psnenbh_li**), then the chance of being in a responding household (complete the household grid and the household questionnaire) at Wave 2 of *Understanding Society* conditional on being issued (reflected in **b_psnenbh_lw**). The weight **b_psnenbh_lw** is then extrapolated to TSMs and PSMs through a weight-share method to create **b_psnenbh_xw**. The detailed procedure for creating these weights as well as cross-sectional individual response weights is described below.

The inclusion weight (**b_psnenbh_li**) was calculated separately for a) Northern Ireland and b) England, Scotland and Wales. For each, it has two components. For Northern Ireland, the first component consists of the BHPS Wave 11 cross-sectional weight, as this is the wave at which Northern Ireland first entered the BHPS. This component encompasses a design weight, post-stratification and an adjustment for Wave 11 nonresponse. The second component is derived from a model of the propensity to be issued at *Understanding Society* Wave 2 conditional on being

enumerated in BHPS Wave 11. This therefore adjusts for all the stages of dropout between BHPS Wave 11 in 2001 and *Understanding Society* Wave 2 in 2010. Model covariates were taken from the Wave 11 household grid and household questionnaire. This propensity was modelled as a single step from 2001 to 2010 because across-wave response patterns varied greatly between the sample members. There is no single BHPS wave since Wave 11 at which all the Northern Ireland sample members (of those issued to *Understanding Society*) responded and therefore no other survey instrument that can provide model covariates for all relevant sample members.

Similarly, for England, Scotland and Wales the first component consists of the BHPS Wave 9 longitudinal weight, as this is the wave at which the Scotland and Wales boost samples were added (so, all of the members of those samples who entered *Understanding Society* were enumerated at that wave, as were the vast majority of members of the original BHPS Wave 1 sample who entered *Understanding Society*). This component therefore encompasses a design weight, Wave 1 post-stratification and adjustments for nonresponse at each of the Waves 1 to 9 of the BHPS. The second component is derived from a model of the propensity to be issued at *Understanding Society* Wave 2, conditional on being enumerated in BHPS Wave 9. This adjusts for all the stages of dropout between BHPS Wave 9 in 1999 and *Understanding Society* Wave 2 in 2010. Model covariates were taken from the Wave 9 household grid and household questionnaire.

BHPS OSM newborns since Wave 9 (England, Scotland or Wales) or Wave 11 (Northern Ireland) whose parents are both OSMs were then assigned a base weight equal to the smaller BHPS inclusion weight of their (OSM) parents in the child's 2010 (issued to *Understanding Society*) household. This reflects the idea that the probability of the child entering the *Understanding Society* sample equals the probability of at least one of his or her parents entering the sample, which in turn is equal to (or greater than) the probability of the parent who has the greatest probability of entering the sample. BHPS OSM newborns born to one OSM parent and one TSM parent were assigned a base weight equal to half of the OSM parent's weight in the child's 2010 (issued to *Understanding Society*) household. The division by two reflects the idea that these newborns had double the chance of becoming BHPS OSMs, relative to people born to both OSM parents, as they would have been included had either their mother's or father's 1991 household been sampled. For newborns observed with a single parent in a household in the first wave after their birth, the weight given was the parent's weight. This reflects a close to zero likelihood for the baby to be sampled via the other parent.

The adjustment for household nonresponse at Wave 2 was derived from a model of enumeration at Wave 2 conditional on entering the *Understanding Society* sample (i.e. being issued to the field for Wave 2), in which covariates came from the Wave 9 household instruments for England, Scotland and Wales and the Wave 11 household instruments for Northern Ireland. The weight which reflects the chance of a BHPS OSM of being selected into the BHPS, to be issued into *Understanding Society* and to be enumerated at Wave 2 of *Understanding Society* is the BHPS-2010 longitudinal enumerated person weight (**b_psnenbh_lw**).

Finally, the BHPS cross-sectional enumeration weight (**b_psnenbh_xw**) was created through a weight-share method by sharing the BHPS-2010 longitudinal enumerated person weight to TSMs and PSMs.

The BHPS cross-sectional weights for main, proxy or telephone interview respondents (**b_indpxbh_xw**), main interview respondents (**b_indinbh_xw**) and self-completion respondents (adults (**b_indscbh_xw**) and youth (**b_ythscbh_xw**) each consist of the cross-sectional individual enumerated weight with an additional adjustment for nonresponse to the relevant instrument conditional on household response. These adjustments were based on logistic regression models with both individual-level and household-level covariates taken from responses to the *Understanding Society* Wave 2 household grid and household questionnaire.

The BHPS cross-sectional household weight (**b_hhdenbh_xw**) is set equal to the minimum cross-sectional person enumerated weight (**b_psnenbh_xw**) amongst adults in the household.

3.3.3.6. Combined Sample (BHPS, GPS and EMBS) Weights

We provide both cross-sectional weights (starting at Wave 2) and longitudinal weights (starting at Wave 3) for combined analysis of respondents in the three samples (BHPS, GPS and EMB). The weights are based on the inclusion enumeration weight, which accounts for combined probabilities of being selected in any of the continuing samples of BHPS, GPS and EMBS at the time each was selected, and continuously being enumerated up to and including Wave 2.

We first explain the calculation of the inclusion enumeration weight, the development of cross-sectional weights for Wave 2 based on it, then the calculation of longitudinal weights, and finally the cross-sectional weights that are created from Wave 3 onward.

3.3.3.7. BHPS, GPS and EMBS Inclusion Enumeration Weight

To combine samples from the BHPS, GPS and EMBS components, we calculate the joint probability of each respondent being selected through each sample and to continue being enumerated up to and including Wave 2. Specifically, the following samples are combined:

Sample 1: BHPS 1991 (those living in England, Scotland and Wales)

Sample 2: BHPS 1999 (those living in Scotland and Wales)

Sample 3: BHPS 2001 (those living in Northern Ireland)

Sample 4: GPS 2009-2010 (those living in England, Scotland, Wales and Northern Ireland)

Sample 5: EMBS 2009-2010 (those living in HDAs in England, Scotland and Wales)

Each respondent therefore had up to five chances to be selected into *Understanding Society*, depending on where they lived at the time that each sample was selected. To reflect this, for each person (selected into any of the five samples) we calculate their probability of being sampled in each of the samples above (p_{jk}, j = 1, … , 5) and add these together to derive the overall inclusion probability ($p_{\bullet k}$):

$$p_{\bullet k} = \sum_{j=1}^{5} p_{jk}$$

The probability of respondent *k* entering UKHLS through sample *j*, p_{jk}, is calculated as:

$p_{jk} = s_{jk} \times h_{jk} \times i_{jk} \times v_{jk}$, where

- s_{jk} is the selection probability of respondent *k* in sample *j*;
- h_{jk} is the average household response rate at Wave 1 for sample *j* in the country of residence of respondent *k* at the time of sample selection;
- i_{jk} is the average individual continuous enumeration probability (rate) between Wave 1 for sample *j* and Wave 2 of *Understanding Society*, specific to respondent *k*'s country of residence at the time of selection of sample *j*. This is in effect to the inverse of the average individual attrition rate.
- v_{jk} is the individual-specific variation from the mean continuous enumeration probability. This is explained in detail below.

For the sample in which the individual was selected we know the selection probability and modelled response probabilities (these can be obtained from the available weights, and the method is described below). These probabilities will be called real. For the samples that the individual could have been selected through but wasn't, we can infer each of the components. Such probabilities are called inferred. The inference method is described below.

Selection probability (s_{jk})

For *j* = 1,2,3, s_{jk} for respondents known or inferred to have been resident in relevant country at the time of selection is equal to the number of eligible households divided by the number of residential households in the population at that time. Such probabilities are calculated separately for each country (England, Wales, Scotland and Northern Ireland). The population information was obtained from Census 1991 for the 1991 sample (Office for National Statistics 1991), and from Census 2001 for samples selected in 1999 and 2001 (Office for National Statistics 2003). For respondents known or inferred not to have been resident in relevant country at the time of selection, $s_{jk} = 0$.

For respondents actually selected into the GPS or EMBS, the real probabilities are simply the inverse of the design weight (**a_hhdenus_xd**) the calculation for which is described above. Two inferred probabilities are calculated for respondents actually selected into the BHPS samples: the probability of selection through GPS (s_{4k}) and the probability of selection through EMBS (s_{5k}). s_{4k} is equal to the number of eligible residential household selected through GPS divided by the number of households in the population according to Census 2011. These probabilities are country-specific.

s_{5k} depends on the postcode sector of residence in 2009-10 (as described in Berthoud et al, 2009), and on the ethnic composition of the household and the selected person. For BHPS sample members we assigned the EMBS postcode sector probability of the postcode of their address at Wave 2 of *Understanding Society* (2010). If the sector was in a LDA, then $s_{5k} = 0$.

Next, for BHPS sample members in high density (ethnic minority) areas we need information on the ethnic group composition of the household. We used information from variables **race** and **racel** as recorded in data file **xwavedata** at the time of

BHPS Wave 18, and **b_racel** as recorded at Wave 2 of *Understanding Society*. For children under 16, for which this information was missing, we inferred it from their biological parents. The information collected in the BHPS does not perfectly match the ethnic group classification used for EMBS selection, but after consulting experts on migration research we achieved a reasonable fit to the EMBS classification. While the selection into EMBS depended on whether a person belongs to the group of interest or not (see section on EMBS), the selection probability depends on the household ethnic group composition at the time of selection (considered 2010 for this purpose). If the household did not have any ethnic minority member of interest, we set $s_{5k} = 0$. If the household had some ethnic minority members of interest, the probability was calculated as the product of postcode sector selection probability and the largest screening probability among ethnic group members. All ethnic minority members were assigned this selection probability, and members of any other groups (either British or those not of selected in EMBS) were assigned the value of zero.

Additionally, postcode sectors from which the Bangladeshi boost was selected were identified and those of Bangladeshi origin were assigned a probability reflecting the additional chance of being selected into Bangladeshi boost, see Berthoud, Fumagalli et al. (2009).

Average household response rate at Wave 1 *(h_{jk})*

We treat nonresponse correction as having three components, the first of which is the household response probability at the first wave when the sample is selected. Here real and inferred probabilities are calculated in the same way. We divide the number of responding households by the number of eligible residential households in the sample – these are country and time specific. For the EMBS the household response rate was calculated using the design weight to account for unknown eligibility.

Average individual continuous enumeration probability (i_{jk})

The second nonresponse component is an average individual continuous enumeration probability. This probability is calculated as the ratio of the number of people who were enumerated continuously in all waves since selection up to and including Wave 2 of *Understanding Society* to the number of people who were enumerated in the wave of selection minus those known to have become ineligible (deceased or out of scope). In other words this is the response rate (where nonresponse is defined as missing at least one wave).

Individual-specific variation from the mean continuous enumeration probability (v_{jk})

The third nonresponse component reflects the variability of enumeration propensities among different individuals. We obtain it by the following procedure: first, we invert the longitudinal enumeration weight for Wave 2 of *Understanding Society* (**b_psnenus_lw** for the GPS and EMBSs and **b_psnenbh_lw** for the BHPS sample); second, we divide this inversed weight by three probabilities described above: selection probability (s_{jk}), average household probability of respond in Wave 1 (h_{jk}), and average individual enumeration probability (i_{jk}). The remainder is then scaled to a mean of 1.00 within each country and each time point of selection, reflecting the enumeration probability for each person, relative to the average person

selected at the same time in the same country. This is the real unique component available for the time points when the person was actually selected.

To obtain tv_{jk} for the time points when a person could have been selected but wasn't, we used the following procedure. For each country and time point we ran a multiple stepwise regression with v_{jk} as the dependent variable and covariates from the household questionnaire and household grid from Wave 2 of *Understanding Society* as explanatory variables. Note that identical predictors are available for all samples of interest. We can therefore use the regression model to infer v_{jk} for sample members selected through another sample.

Residency

In order to assign a correct probability for each person we need to know where they lived in 1991 (England, Scotland, Wales or other), 1999 (Scotland, Wales or other), 2001 (Northern Ireland or other) and 2009/2010 (postcode sector if England, Scotland and Wales, Northern Ireland, other). Because no interviewing was conducted among BHPS members in 2009, we treat their address at the time of Wave 2 of *Understanding Society* (2010) as if it were their address at the time of selection of GPS and EMBS.

We do not have perfect information for each enumerated respondent on their residency in the four years of interest, but we use all the available information to us to infer residency as closely as possible.

Specifically, the following information is used: the country and year of the respondents' selection; the respondent's report on the most recent change of address; for those not born in Britain, the year of their arrival to Britain the country and year of school or university studies. For adults whose residence remains unknown after using the above information, it is inferred from other household members, if it is consistent.

Note, that for those who were born after the sample was selected, we make use of residency information from their mother. In this way we obtain residency information with five categories (England, Wales, Scotland, Northern Ireland or abroad) for 1991, 1999, 2001 and 2009/2010 time points for each single respondent.

Total probability

The important point is that an estimate of each of the following probabilities is available for each person, regardless of when and where the person was selected:

p_{1k} if resident in England in 1991;

p_{1k} if resident in Scotland in 1991;

p_{1k} if resident in Wales in 1991;

p_{2k} if resident in Scotland in 1999;

p_{2k} if resident in Wales in 1999;

p_{3k} if resident in Northern Ireland in 2001;

p_{4k} if resident in England in 2009/10;

p_{4k} if resident in Scotland in 2009/10;

p_{4k} if resident in Wales in 2009/10;

p_{4k} if resident in Northern Ireland in 2009;

p_{5k} if resident in England in 2009/10;

p_{5k} if resident in Scotland in 2009/10;

p_{5k} if resident in Wales in 2009/10.

For each time point (1991, 1999, 2001, and 2009/2010) the probability is non-zero only for the sample reflecting the respondent's country of residence at that time, and is zero for all other countries for that time point. Thus for a person who has always lived in England the non-zero selection probabilities will be those for England in 1991 and for England for 2009/2010. For those who immigrated to England in 2008, all selection probabilities will be zero except for the probabilities for England in 2009/2010.

The total probability, $p_{\bullet k}$, is therefore the sum of all the above probabilities. It reflects multiple possible ways of selection through different samples and continuous enumeration of a respondent up to and including Wave 2 of *Understanding Society*. The inclusion weight (**b_psnenub_li**) is the inverse of the total probability. This weight is not scaled and will serve as a base weight for all the weights that combine BHPS, GPS and EMBSs. The weight is defined for all OSM respondents enumerated at Wave 2 of *Understanding Society*.

3.3.3.8. *Wave 2* Cross-sectional Weights for Combined Sample (BHPS, GPS and EMBS Components)

Cross-sectional weights are created for Wave 2 of *Understanding Society* for the combined samples of BHPS, GPS and EMB. The first is the cross-sectional individual enumeration weight which is created through the weight share technique where the inclusion weight (**b_psnenub_li**) is weight-shared to TSMs and those OSMs that have missed at least one wave between selection and Wave 2 of *Understanding Society*. The weight share is done in a standard way: the inclusion weight, **b_psnenub_li**, is unaltered if it is non-zero for all household members but the cross-sectional weight is equal to the average inclusion weight for households where at least one member has a value of zero for the inclusion weight. The obtained cross-sectional enumeration individual weight is then scaled to a mean of one (**b_psnenub_xw**).

For each household the lowest **b_psnenub_li** from OSM adults (16 years of age or older) is selected. This reflects the highest probability of enumeration among OSM household adult members. The weight is then scaled to the mean of one (**b_hhdenub_xw**).

Additionally, three adult interview weights are provided: **b_indpxub_xw**, **b_indinub_xw** and **b_indscub_xw**. To create these weights we use the respective BHPS cross-sectional weight (**b_indpxbh_xw**, **b_indinbh_xw** and **b_indscbh_xw**) for the BHPS sample and GPS and EMBS cross-sectional weights (**b_indpxus_xw** **b_indinus_xw** and **b_indscus_xw**) for the GPS and the EMBS. No additional scaling is required because each sample (BHPS vs. GPS+EMBS) contributes in proportion to the combined sample.

3.3.3.9. Longitudinal Weights for Combined Sample (BHPS, GPS and EMBS Components) for *Wave 3* onward

A longitudinal enumeration weight for the BHPS, GPS and EMBS (**c_psnenub_lw**) was created based on the inclusion enumeration weight into Wave 2 (**b_psnenub_li**). Conditional on a nonzero value for **b_psnenub_li**, the enumeration is modelled using logistic regression and predictors from the Wave 2 household questionnaire and household grid. The estimated probabilities of conditional enumeration are inversed, and multiplied by **b_psnenub_li**. Newborns are given their mother's enumeration weight. The weight is then scaled to a mean of one.

Longitudinal response weights for proxy and main interview, main interview, and self-completion interview were also created at Wave 3. All of these are based on longitudinal enumeration in Wave 3 (positive value of **c_psnenub_lw**). For each instrument, a logistic regression is run to predict response in both Waves 2 and 3 with predictors from the Wave 3 household questionnaire and household grid. The models are restricted to adults. The estimated probabilities were then inversed, multiplied by **c_psnenub_lw** and scaled to a mean of one.

From Wave 4 onward, each of these longitudinal weights are based on the equivalent weight from the previous wave, adjusted by the reciprocal of the predicted value from a logistic regression model of wave-on-wave response and scaled to a mean of one. For example, a logistic regression model of enumeration at Wave 4 was based on sample members with a non-zero value of **c_psnenub_lw** and the reciprocal predicted values were multiplied by **c_psnenub_lw** and then scaled to produce **d_psnenub_lw**.

3.3.3.10. Cross-sectional Weights for Waves 3 onward

Starting at Wave 3, cross-sectional weights were created for the combined sample that includes the BHPS, GPS and EMBS components. The cross-sectional enumeration weight (***n*_psnenub_xw**) is created based on the longitudinal enumeration weight (***n*_psnenub_lw**) via the weight-share method. The weight is shared from OSMs with nonzero longitudinal enumeration weights to TSMs and PSMs (except those who were selected at Wave 1), and OSMs with a longitudinal weight of zero.

For the household cross-sectional weight (***n*_hhdenub_xw**) the lowest cross-sectional enumeration weight among adults (***n*_psnenub_xw**) is selected. The weight is then scaled to a mean of one.

Cross-sectional weights for proxy and main (***n*_indpxub_xw**), main (***n*_indinub_xw**) and self-completion (***n*_indscub_xw**) are created based on the Wave *n* cross-sectional enumeration weight. Three logistic regressions are run to predict the response for the relevant instrument with predictors from the household grid and household questionnaire in Wave *n*, conditional on enumeration in Wave *n*. The models are restricted to adults. The estimated probabilities are inversed, multiplied by ***n*_psnenub_xw** and are scaled to a mean of one.

The cross-sectional weight for youth (***n*_ythscub_xw**) is also created based on enumeration at the same Wave *n* (***n*_psnenub_xw**). Among eligible youth members (aged 10 to 15) a logistic regression is run with predictors from the household grid and household questionnaire from Wave *n* to predict response to the youth

questionnaire. The inverse of the probability is multiplied by ***n*_psnenub_xw** and is scaled to a mean of one.

3.3.3.11. Longitudinal Weights for Combined Sample (BHPS, GPS, EMBS and IEMBS Components) for *Wave 6* onward

A longitudinal enumeration weight for the BHPS, GPS, EMBS and IEMBS (**g_psnenub_lw**) was based on the inclusion enumeration weight into Wave 6 (**f_psnenub_li**), and was created in a similar fashion to the longitudinal enumeration weight for the BHPS, GPS and EMBS combined sample (**c_psnenub_lw**).

Longitudinal response weights for proxy and main interview (**g_indpxui_lw**), main interview (**g_indinui_lw**), and self-completion interview (**g_indscui_lw**) were also created at Wave 7. All of these were based on longitudinal enumeration in Wave 7 (positive value of **g_psnenub_lw**), and were created in a similar fashion to the respective longitudinal weights at Wave 3 for the BHPS, GPS and EMBS combined sample.

Longitudinal weights ***n*_psnenui_lw** , ***n*_indpxui_lw, *n*_indinui_lw, *n*_indscui_lw** starting from Wave 8 onwards are calculated in a similar process to the respective weights for the BHPS, GPS and EMBS combined sample.

3.3.3.12. Cross-sectional Weights for *Wave 6* onward

In addition to the cross-sectional weights for the BHPS, GPS and EMBS combined sample, starting at Wave 6 we provide cross-sectional weights for the BHPS, GPS, EMBS and IEMBS combined sample. The cross-sectional enumeration weight (*n*_psnenui_xw) is based on the longitudinal enumeration weight (*n*_indenui_lw) and is created via the weight-share method in a similar process to creating the cross-sectional weights for the BHPS, GPS and EMBS combined sample (*n*_psnenub_xw). The non-eligible for the weight-share TSMs include not only TSMs selected at Wave 1 as part of EMBS, but also those TSMs that were selected at Wave 6 as part of IEMBS.

The other cross-sectional weights including the household cross-sectional weight (*n*_hhdenui_xw), the cross-sectional weights for proxy and main (*n*_indpxui_xw), main (*n*_indinui_xw) and self-completion (*n*_indscui_xw) interviews, and the cross-sectional weight for youth (*n*_ythscui_xw) are created using the same process as the respective cross-sectional weights for BHPS, GPS and EMBS combined sample. All of these weights are based on the cross-sectional enumerated person weight for the BHPS, GPS, EMBS and IEMBS combined sample (*n*_psnenui_xw).

3.3.3.13. Longitudinal weights to analyse biomarker data

Each wave has four weights for longitudinal analysis of adult full interviews together with the data from nurse visit and blood measures. In the designated user guide for this data collection Benzeval, Davillas et al. (2014) describe the development of cross-sectional weights to analyse biomarker data and the longitudinal biomarker weights for Wave 3. In each subsequent wave we create ***n*_indns91_lw** (for longitudinal analysis of BHPS data since 1991 and including wave *n* with the nurse visit data), ***n*_indnsub_lw** (for longitudinal analysis of BHPS, EMBS and GPS data since Wave 2 up until wave *n* with the nurse visit data), ***n*_indbd91_lw** (for longitudinal analysis of BHPS data since 1991 and including wave *n* with the blood measures) and ***n*_indbdub_lw** (for longitudinal analysis of BHPS, EMBS and GPS

data since Wave 2 up until wave *n* with the blood measures). All of these weights are created based on the previous wave's corresponding weight with a nonresponse adjustment for nonresponse to the full adult interview. Thus, the correction is a reciprocal of the predicted response from a stepwise logistic regression using predictors from the previous wave household grid and questionnaire, and from the adult full questionnaire. This correction is then multiplied by a corresponding previous wave weight. The obtained weight is then truncated based on the design effect and scaled to the mean of one.

An exception to the above general way of creating weights is the longitudinal weight ***w*_indbd91_lw** for the original 1991 BHPS sample for blood measures in some waves (e.g. Wave 4 and Wave 5). This is because the nonresponse rate in Wave *n* conditional on *n-1* is often extremely low; hence no predictors of nonresponse are found to be statistically significant. Thus, the Wave *n* ***n*_indbd91_lw** is proportional to the ***(n-1)*_indbd91_lw** weight. The combined nonresponse between Waves *n-1* and *n+1* in then corrected in Wave n+1. For this the base weight was ***(n-1)*_indbd91_lw** and the combined nonresponse over two waves is predicted using stepwise logistic regression with predictors from Wave n-1, thus creating **(*n+1)*_indbd91_lw**.

3.4. DERIVED INCOME VARIABLES

3.4.1. OVERVIEW

Understanding Society collects detailed information each wave on personal income. All individuals aged 16 or more are asked to report:

- wages,
- self-employment earnings,
- second job earnings,
- interest and dividends,
- pensions (National Insurance/state retirement pension, pension from a previous employer, pension from a spouse's previous employer, private pension/annuity, widow's or war widow's pension, widowed mother's allowance or widowed pension),
- benefits (severe disablement allowance, disability living allowance, war disablement pension, attendance allowance, carer's allowance, incapacity benefit, income support, job seeker's allowance, national insurance credits, child benefit, child tax credit, working tax credit, maternity allowance, housing benefit, council tax benefit, foster allowance/guardian allowance/rent rebate, rate rebate, employment and support allowance, respond to work credit, sickness and accident insurance, in-work credit for lone parents and pension credit) and
- other income sources (educational grant, trade union and friendly society payment, maintenance or alimony, payments from a family member not living together, amount for rent from boarders or lodgers, rent from any other property).

These personal income variables can be summed to obtain the total personal income. Total household income can be computed from the personal total incomes of all household members. We provide these derived variables as well as net income

estimates, discussed below, as part of the released data files. We also provide derived variables relating to housing costs so that measures of income after housing costs can be computed.

3.4.2. IMPUTATION OF INCOME VARIABLES

Some of the income components can be missing. More precisely there can be three types of missing cases:

- item nonresponse when individuals respond to the individual questionnaire but do not answer to some or all the questions on income components;
- individual nonresponse when individuals fail to respond to the individual questionnaire;
- household nonresponse when there is neither a household nor the individual questionnaire response.

For example, at Wave 1 we have 59,466 individuals for whom at least the household questionnaire is available, and, among these individuals, 80.3% provided a personal interview, 5.5% have a proxy interview, whereas 14.2% had neither a proxy nor a personal interview. The item nonresponse rate for individuals who provided an individual questionnaire varies across income variables. It goes from a maximum of about 50% for self-employment earnings to zero for some of the benefit variables, and it is generally below 20% for the remaining income variables.

3.4.2.1. What Do We Impute?

In *Understanding Society* we do not impute income variables for non-responding households. Responding households are households for which the household questionnaire and information on the household composition/structure (household grid module) are available. We suggest that the user take account of household nonresponse via weighted estimates, see Section 3.3).

For individuals who respond to the individual questionnaire but do not provide answers to all income questions (item nonresponse), we impute the following personal income variables: wages, self-employment earnings, second job earnings, interests and dividends, pensions, benefits and other income sources.

For individuals for whom a proxy questionnaire is available, we include in the data set imputations for total earnings and total income whenever missing. The proxy questionnaire is a short version of the individual questionnaire with questions on total earnings and total income as well as other variables. Our imputation procedures also impute the individual income components for proxy respondents. These are used to calculate total household income components but individual level values are not included in the data set.

Finally, for individuals in responding households for whom neither the personal nor the proxy questionnaire is available, our imputation procedures also impute the individual income components. These are used to calculate total household income and components of household income by source but individual level values are not included in the data set.

Based on these imputations we can compute total personal and household income for all individuals belonging to responding households.

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 this takes 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**.

Receipt of benefits and some other sources is recorded in a separate data file, i.e., ***w*_income**. 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. These are summed and imputed as such, and the imputed values are in the variable ***w*_frmnthimp_dv**. As a consequence, the variable ***w*_frmnthimp_dv** for the first income receipt from a given source is equal to the total value of all receipts from that source, while it is set to zero for the second and subsequent receipt. Some income sources may be reported by more than one member of the household. In order to avoid double counting the derived variable ***w*_frjtkeep_dv** identifies which of these will be included in income totals.

3.4.2.2. Imputation Procedures

Missing income values in *Understanding Society* are replaced using a combination of cross-sectional and longitudinal imputation methods. There are two steps to the *Understanding Society* imputation procedure. The first replaces missing values using cross sectional imputation methods (with some exceptions where we make use of longitudinal information (see carryover below)). A second step then replaces the first stage imputes using the longitudinal imputation method of Little and Su (1989). The overall approach is based on that used for the Australian household panel, (HILDA), described in Hayes and Watson (2009). The various imputation methods used are described in more detail below. With some exceptions noted below, respondents to the full individual questionnaire, respondents by proxy and within-household non-respondents are all included together in the imputation models.

Cross sectional imputation methods

Cross-sectional imputation is carried out year by year through a range of parametric, semi-parametric and non-parametric methods. Parametric methods are: linear regression (for continuous variables), interval regression (for continuous censored variables), logistic regression (for binary variables), ordered logistic regression (for ordered variables), multinomial logistic regression (for non-ordered categorical variables). The semi-parametric and non-parametric methods are, respectively, predictive mean matching (PMM) and hot-deck imputation.

Parametric methods and PMM are used to impute wages, self-employment earnings, second job earnings, interests and dividends, plus their predictors for responding individuals. For responding individuals, wages, self-employment earnings, second job earnings, interests and dividends, and their predictors are imputed jointly using chained equations (ICE). Hot-deck is used to impute income sources for proxies and non-respondents (missing values in the variables defining the categories are set equal to their median).

All variables are imputed as reported except for wages and self-employment income, where we convert amounts reported net to gross where gross is not reported, using a deterministic model based on the tax and national insurance system. In computing total personal income, it is assumed that all other sources are reported gross, or are

not subject to taxation. Net income estimates are also included in the data set (see Section 3.4.3).

In what follows, we outline briefly the characteristics of the main cross-sectional imputation methods used.

Chained equations (ICE) is a multivariate imputation method used to impute a set of variables jointly. We used it to impute the main income variables for respondents plus their predictors. ICE allows for interdependence between income and auxiliary variables by considering univariate models estimated separately and sequentially through stochastic imputation, see van Buuren, Boshuizen et al. (1999) and Ragunathan, Lepkowski et al. (2001). This method has been already used in some major household panel surveys such as the ECHP.

The ICE starts by considering the following recursive (triangular) system of imputation equations,

$$\begin{cases} Y_1 = \alpha_{10} + X\beta_1 + u_1 \\ Y_2 = \alpha_{20} + X\beta_2 + \alpha_{21}Y_1 + u_2 \\ Y_3 = \alpha_{30} + X\beta_3 + \alpha_{31}Y_1 + \alpha_{32}Y_2 + u_3 \\ \vdots \\ Y_k = \alpha_{k0} + X\beta_k + \alpha_{k1}Y_1 + \alpha_{k2}Y_2 + \cdots + \alpha_{kk-1}Y_{k-1} + u_k \end{cases}$$

Here, Y1,Y2,…,Yk are the income and auxiliary variables to be imputed ordered from the one with the smallest percentage of missing values, Y1, to the one with the largest percentage of missing values Yk, X is a set of auxiliary variables observed for all individuals, α's and β's are parameters and u1,u2,…,uk are random errors. Such a recursive system allows us to carry out the imputation separately for each variable and sequentially. The sequential procedure is given by the following steps:

- estimation of the first equation and imputation of the missing values for Y1,
- estimation of the second equation using the imputed values to replace the missing values of Y1, and imputation of Y2,
- repetition of estimation and imputation steps sequentially for each of the following equations until when all k variables, Y1,Y2,…,Yk have been imputed.

We use stochastic imputation, that is, we draw the imputed values from the posterior predictive distribution of the variable to be imputed, conditional to the observed data. For more details about stochastic imputation we refer to Rubin (1987), Schafer (1997), and Kenward and Carpenter (2007).

This sequential estimation is consistent only if the recursive system is valid. Since this is not necessarily a valid assumption, ICE uses the imputed values produced using the above recursive system as starting values in an iterative imputation process. In other words, the starting values are used to begin a new cycle of imputations where each equation is estimated sequentially, but this time using as explanatory variables both X and all the imputed variables Y1,Y2,…,Yk excluding the one used as dependent variable. At the end of this new cycle, a set of new imputed variables is produced and used to begin a further new cycle of imputations. These cycles of imputations are repeated until convergence. Notice that in practice some of the variables will exclude certain of the Xs and Ys variables in the imputations because it does not always make sense to use all variables as predictors.

Predictive mean matching (PMM) is used to impute benefits, pensions and other incomes. For a given variable, PMM replaces missing values with observed values from a donor, i.e., a respondent with non-missing information on the variable of interest. This is done in four steps:

a) regression models for the variable to be imputed are estimated
b) fitted values are produced
c) records with missing information (recipients) are matched to donors based on the fitted values computed in b),
d) missing values are replaced with observed values from donors.

See also Little (1988).

Hot-deck (HD). For individuals with missing information, the hot-deck method identifies suitable donors within imputation classes. Characteristics reported in the data associated with the missing information are used to define imputation classes.

In *Understanding Society*, the hot-deck method is used to impute missing information for proxies and non-respondents. For proxies, where available, imputation classes are defined by reported bands on income and earnings and limited covariates. For non-respondents and proxies where income bands are missing, a richer set of covariates are used to define imputation classes including: employment and benefit carryovers, age, education, sex, ethnicity, housing tenure, marital status, durable good ownership, whether a parent and number of bedrooms. Once a suitable donor is identified, information on all income sources is carried over from the donor.

Longitudinal imputation methods

Methods for longitudinal imputation are used to take into account longitudinal patterns in the data. The longitudinal methods we use are the “Population Carryover”-method and the “Little and Su”-method.

Population Carryover (PC-method). The PC-method uses data from adjoining waves to replace missing wave information. With only one adjoining wave of non-missing data, the information is carried-over with probability one. When two waves of adjoining information are available, the information carried-over is chosen based on proportions reported in the non-missing population.

In *Understanding Society*, the PC-method is used to impute employment status and benefit eligibility for non-respondents and proxies. These variables are then used as inputs into a hot-deck procedure (see above).

Little and Su (LS-method). The LS-method imputes missing values using a multiplicative model (see Little and Su 1989). The final imputation is the product of three terms: a trend effect (across waves), the recipient’s departure from the trend, and a residual effect donated from another respondent with complete information for the corresponding income component.

In *Understanding Society*, a modified version of the LS-method is implemented. When identifying a donor for the residual effect, rather than using the non-missing information only, we additionally make use of information imputed from the cross-sectional methods described above. In this way, the LS-method forms the final step in the *Understanding Society* imputation process, where imputes from the other methods form inputs into the LS-method. Missing income sources imputed as

inapplicable using the cross-sectional methods do not receive an LS-method impute, neither do respondents applicable for an income source in only one wave.

3.4.2.3. Specification of Cross-Sectional Imputation

The imputation of earnings (wages and self-employment earnings from the first job and earnings from the second job) and investment income in the individual questionnaire is performed considering a separate equation for each of the income components.

With variables for which we have point information (a single value), we use either log linear or predictive mean matching models. For those variables where we have bracketed values rather than point information (for example in the case of dividends and interests) or when we have *a priori* information which allows us to bound the missing income variable, we use interval regression. The type of regression models used to impute missing explanatory variables depends on the level of measurement. That is, we use log linear regression for continuous variables, and binary, ordered and multinomial logit models, respectively, for dummy, ordinal and unordered categorical variables. The explanatory variables are a set of characteristics collected in the individual (personal) or household questionnaires. The specification of the models varies by income variable but it generally includes most of the following variables:

- personal socio-economic variables (age, sex, self-reported ethnic group, indicator for respondent born in the UK, marital status, education level, general health, current subjective financial situation);
- personal income variables (excluding the one used as the dependent variable);
- lagged income variables (just for Waves 2 and 3)
- household characteristics (number of children in the household, house tenure, house type, household size);
- job characteristics (log number of hours normally worked per week, log number of hours per months in a second job, log years of job tenure, permanent or temporary job, occupation (SOC 2000, 1 digit), number employed at the current job workplace (for employees), number of employees if self-employed, whether is self-employed and hires employees, whether the employment organization is private or not (only for employees), type of ownership if self-employed (sole ownership or partnership), an indicator for whether annual business accounts are prepared for the Inland Revenue for tax purposes if self-employed);
- household variables reflecting economic situation (log amount spent on food from food shops in four weeks prior to interview, log amount spent on food eaten outside the home in four weeks prior to interview, log last year expenditure on domestic fuel (e.g. electricity and gas), number of bedrooms in the house, number of other rooms in the house, Council Tax band);
- Government Office Regions (GOR).

The imputation of the income sources in the income file (pensions and benefits) is performed in a second model where each income source is imputed using as predictors the other income sources of the income file, a set of demographics (age, age squared, number of children, number of children squared, sex, ethnicity, marital

status, GOR), the income sources imputed in the previous stage (earnings for first and second job and investment income), and information on benefits and pensions in the previous year (total value and total number of benefits).

All variables are imputed as reported except for wages and self-employment income, where we convert amounts reported net to gross where gross is not reported, using a deterministic model based on the tax and national insurance system. In computing total personal income, it is assumed that all other sources are reported gross, or are not subject to taxation.

The imputation of the missing income sources in the individual questionnaire permits the computation of total earnings (i.e., the sum of income from the first and second job) and total income (the sum of earnings, plus investment income, pensions and benefits).for all adult non proxy respondents.

3.4.3. NET INCOME ESTIMATES

The data also include estimates for monthly income net of tax and national insurance. We have worked with the DWP to replicate as far as possible the approach they use in developing their Households Below Average Income (HBAI) estimates. There are however some deductions from individual income and some incomes sources which are not available to us from the questionnaire. It is hoped that a later release of the data will provide further estimates of some of these other components. A technical working paper providing more information the derivation of these net income variables is in preparation. The DWP use *Understanding Society* data for the longitudinal component of UK statistics on income dynamics.

Individual income estimates are included in the individual level data files, ***w***_**indresp** and the household-level income measures are included in the household level data files, ***w***_**hhresp**.

At the individual level, the total estimated net monthly income is ***w***_**fimnnet_dv** where “net” refers to net of taxes on earnings and national insurance contributions. It is constructed from the income components described below. The gross monthly income, ***w***_**fimngrs_dv** is also estimated from individual income components described below except the earnings components are gross, that is, before taxes and National Insurance contributions are deducted. The associated imputation flag for both variables is ***w***_**fimngrs_if**.

At the household level, total household gross income is included in the variable ***w***_**fihhmngrs_dv**. This comprises imputed income from proxy and within-household non-respondents. The extent of imputation is indicated by the variable ***w***_**fihhmngrs_if**. The calculation of housing costs (see below) implies that there is housing benefit implicitly reported in the rent information which has not been reported in the individual questionnaire. ***w***_**fihhmngrs1_dv** includes an adjustment for this.

In addition to the summary variables described above the individual level data files also include estimates of the different income components, following the structure used by HBAI. 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***), which is gross pay in second job (***w_j2pay_dv***), less estimated tax and national insurance.

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)".

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".

Component 5: 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 - less estimated tax". 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.

Component 6: 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".

Component 7: 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 (offset against council tax)", [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).

At the household level, the following income variables are available:

***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**).

The variables ***w_*fihhmninv_dv, *w_*fihhmnlabgrs_dv, *w_*fihhmnlabnet_dv, *w_*fihhmnmisc_dv, *w_*fihhmnpen_dv, and *w_*fihhmnprben_dv** provide

individual components of income, summed over all household members including proxies and within-household non-respondents, making use of individual income component imputations.

***w*_fihhmnnet3_dv** is equal to ***w*_hhnetinc1** less council tax liability. This is made available only with the *Understanding Society* Special Licence ***w*_hhresp** data files, see UKDS study number SN6931.

***w*_fihhmnnet4_dv** is equal to ***w*_hhnetinc1** less council tax liability and also adjusted for housing benefit as reported in the household questionnaire as distinct from that reported in the individual questionnaire (using ***w*_hbadjust_dv**). This is made available only with the *Understanding Society* Special Licence ***w*_hhresp** data files, see UKDS study number SN6931.

Council tax liability, for most people, is equal to their estimated council tax, ***w*_ficountax_dv** (see below). Some people receive council tax benefit to help them pay their council tax (see item 23 in component 7 described above). For these people council tax liability equals their estimated council tax minus their council tax benefit.

***w*_ficountax_dv** is the estimated council tax. It is estimated from council tax band and local authority district (GB). Estimates for Northern Ireland rate charges have not been included. This variable is constructed at the household level and, given its sensitive nature, included only with the *Understanding Society* Special Licence data, see UKDS study number SN6931.

Table 44 lists the components of net income provided with our data.

Table 44: Components of net income variables on *Understanding Society*

Personal monthly income	
*w*_fimngrs_dv	Gross monthly personal income gross (imputed)
*w*_fimnnet_dv	Net monthly personal income (imputed), no taxes deducted other than taxes on earnings
*w*_fimngrs_if	Imputation flag *w*_fimngrs_dv and *w*_fimnnet_dv
Components of (imputed) personal gross monthly income: *w*_fimngrs_dv	
*w*_fiyrinvinc_dv	Gross **annual** income from savings and investments
*w*_fibenothr_dv	Gross monthly income from benefits and other sources
*w*_fimnlabgrs_dv	Gross monthly labour income
Components of (imputed) personal net monthly income: *w*_fimnnet_dv	
*w*_fimnlabnet_dv	Net monthly personal labour income (imputed)
*w*_fimnmisc_dv	Monthly personal miscellaneous income (imputed)
*w*_fimnprben_dv	Monthly personal private benefit income (imputed)
*w*_fimninvnet_dv	Net monthly personal investment income (imputed)
*w*_fimnpen_dv	Monthly personal pension income (imputed)
*w*_fimnsben_dv	Monthly personal social benefit income (imputed)
Components of (imputed) personal net monthly labour income: *w*_fimnlabgrs_dv	
*w*_paynu_dv	Net monthly personal earnings from main job (imputed); same as *w*_paynu_dv
*w*_seearnnet_dv	Net monthly personal self-employment income (imputed); *w*_seearnnet_dv

*w*_j2paynet_dv	Net monthly personal earnings from second job (imputed); j2pay_dv (gross monthly earnings from second job, imputed) MINUS taxes and national insurance contributions
Household monthly income	
*w*_fihhmnnet1_dv	Net monthly household income (imputed), no taxes deducted other than taxes on earnings
*w*_fihhmninv_dv	Sum of monthly amount of investment income received by all household members.
*w*_fihhmnlabgrs_dv	Sum of total personal monthly income from labour income received by all household members.
*w*_fihhmnlabnet_dv	Sum of total personal monthly net income from labour income received by all household members.
*w*_fihhmnmisc_dv	Sum of monthly amount miscellaneous income received by all household members.
*w*_fihhmnpen_dv	Sum of monthly amount of pension income received by all household members.
*w*_fihhmnprben_dv	Sum of monthly amount of private benefit income received by all household members.
*w*_fihhmnsben_dv	Sum of monthly amount of personal social benefit income received by all household members.
w_ficountax_dv	Council tax (estimated)
w_fihhmnnet3_dv	Net monthly household income (imputed), minus taxes on earnings, national insurance contributions and council tax liability
w_fihhmnnet4_dv	Net monthly household income (imputed), minus taxes on earnings, national insurance contributions and council tax liability, adjusted for housing benefit in gross rent

Notes: Personal incomes stored in data file ***w*_indresp**. Household incomes and estimated council tax stored in data file ***w*_hhresp**. Variables in italics are available with the Special Licence version of the data, see UKDS study SN6931.

3.4.4. HOUSING COSTS ESTIMATES

We provide derived variables for total housing costs including imputations in order to allow computation of income after housing cost measures. Derived housing cost variables cover renters and those paying mortgages.

For renters we have created the following:

> ***w*_rent_dv** is the computed monthly net rent paid, using ***w*_rent** and ***w*_rent_wc**. This does not include imputations.
>
> ***w*_rentg_dv** is the computed monthly gross rent including any housing benefit received. It is equal to ***w*_rent_dv** where no housing benefit is received. Missing values are imputed, and where the participant reports 100% housing benefit 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*_rentg_dv**.
>
> ***w*_hbadjust_dv** is an adjustment for total household income where housing benefit is implicitly reported in the difference between gross and net rent, but is not reported in the individual questionnaire.

For those paying mortgages we provide the following:

***w*_xpmg_dv** is monthly total mortgage payments including imputation for missing data on ***w*_xpmg**. 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.

And the above derived variables for rent and mortgages are combined in the following variables:

***w*_houscost1_dv** is total housing costs including capital repayments, i.e. ***w*_rentg_dv** plus ***w*_xpmg_dv. *w*_houscost2_dv** excludes capital repayments, i.e ***w*_rentg_dv** plus ***w*_xpmgint_dv.**

In the imputation of rent and mortgage payment it is assumed that variations over time are small and where other reports at the same address are available, missing values are set equal to the median of these reports. Where no report at that address is available a single value is imputed on the basis of characteristics of the accommodation and household and applied to all relevant waves.

3.4.5. TOP-CODING OF INCOME AND INVESTMENT VARIABLES

Reported and imputed income and investment amounts have been top-coded in order to prevent disclosure. 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 for gross income and an 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 new 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 applied is via UKDS study number SN6931.

4. Further Notes For Analysts

4.1. Not Using Weights

Note, that 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.

If no weighting is used, an analysis of *Understanding Society* data assumes:

- that all estimates of interest are the same in Northern Ireland as in the rest of the UK;
- that people of ethnic minority origin are the same as British;
- that recent immigrants to UK do not differ from people who stay in the country longer;
- that people who live at an address with more than three dwellings or more than three households are the same as those who don't;
- that people who responded at Wave 1 are the same with respect to your estimates as those who did not; that people who continued to respond at later waves are the same as those who did not; and that people who responded to each particular instrument used in the analysis (individual interview, self-completion questionnaire etc.) are the same as those who did not, see Lynn, Burton et al. (2012).

An unweighted analysis of the former-BHPS sample assumes:

- that all estimates of interest are the same in each of England, Scotland, Wales and Northern Ireland;
- that people who live at an address with more than three dwellings or more than three households are the same as those who don't;
- that people who responded at Wave 2 of *Understanding Society* in 2010 are the same with respect to your estimates as those who may have become non-respondents at any time since Wave 1 of BHPS in 1991;
- that people who keep responding in later waves of *Understanding Society* are the same as 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.

4.2. Using Self-completion Variables

Analysts should remember to consult the section on specifying the complex sampling variables from Section 3.2.7 and on weighting from Section 3.3.

Analysts who draw on information collected in the self-completion interview are advised to use the associated self-completion weights, see Section 3.3.

For convenience, a great deal of self-completion variables can be identified through the prefix "*w*_sc". Self-completion variables are listed under the Index Term "Self-completion variables" in the online data documentation.

Note that for respondents interviewed online (from Wave 7 onward) all questions are technically answered in a self-completion mode. Nevertheless, variables in the dataset follow the CAPI naming convention and the weights should be selected based on CAPI questionnaire conventions (i.e., use "px" for proxy or "in" for full main questionnaire or "sc" for self-completion part in CAPI questionnaire).

4.3. USING THE IMMIGRANT AND ETHNIC BOOST SAMPLES

Data from all samples are provided together in the same data files. In most cases we recommend using the data from the different samples together to avoid coverage error (Section 3.3). There are some exceptions – see Sections 4.4 and 4.5. To identify the sample from which an observation is drawn please use the variables ***w*_hhorig** and ***w*_memorig**. Note it is ok to use sub-samples based on different characteristics (such as sex, age, ethnic group, religion, employment status etc.) as long as the observations are taken from across all the design samples.

In Wave 6, a few questions were asked of the IEMBS only. These questions will be asked of the rest of the sample (as long as they satisfy the eligibility rule for that question) in Waves 7 and 9. It is not recommended that users analyse these questions in Wave 6 using only the IEMBS, without very clearly stating the coverage error that would apply to the results. As weights are not available for this sample only, only unweighted analysis can be performed with its accompanying issues of lack of generalizability (see Section 3.3).

4.4. USING THE BHPS

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).

The questionnaires used for the two samples are the same. There are, however, a few differences in the data collected. One important issue is that the date of previous interview for GPS sample members who were interviewed at the previous was approximately 12 months earlier, while for the former BHPS sample the gap was between 13 and 27 months for sample members interviewed at Wave 18 of BHPS. This means that the reference period for the history of events since the last interview will be longer for the BHPS sample. This variation in the reference period within a UKHLS wave applies only to Wave 2.

Both samples can be used for cross-sectional and longitudinal analyses. For the appropriate weights, select from Table 35 through Table 42. Note that

Table 36 through Table 41 apply to analysis at the individual level.

To use the long-run of data collected from BHPS sample members, users have two options. The first option is to use the harmonised BHPS, which is included in the Understanding Society data release. The second option is to use the stand-alone BHPS (SN5151).

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). The data files also include the UKHLS-format household identifier variable (i.e., stem name **hidp** preceded by the wave prefix).

There are advantages in using the harmonised BHPS files: all variables that have an equivalent in both studies have been renamed so they have the UKHLS name and efforts have been made to assure that the information content of variables with the same name is identical. If it was not, the BHPS variable received the suffix _*bh*. The **xwavedat** files from both studies have been merged and all information that is available for cases from both samples has been harmonised.

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 (Fumagalli, Knies et al. 2017).

In matching to earlier waves of stand-alone BHPS data (SN5151) it is important to be aware that variable names in the BHPS data set 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
- derived variables, imputation flags, weights and other special variables are not distinguished by _dv or _if suffixes.
- While the great majority of BHPS sample cases who were interviewed in *Understanding Society* Wave 2 were previously interviewed at Wave 18 in 2008-9, there are a number who were last interviewed at an earlier wave. Information about the response status of BHPS sample members at each of the 18 waves is contained in the BHPS file **xwaveid**.
- The BHPS data set also contains a file called **xwavedat**, which contains the values for stable variables (e.g., ethnic group, parent social class etc.). Because of some differences in variable definition, this information has not been copied across to the new *Understanding Society* file, also called **xwavedat**. However in most cases values of these variables can be obtained by matching to the BHPS file.

However, most questionnaire variables which are carried in both surveys will have the same main variable name, though with a different wave prefix. Since the last wave of BHPS was Wave 18, the wave prefix is “r”. Thus if we wished to match Wave 2 work status (**b_jbstat**) on the file **b_indresp** to previous wave values, for the GPS sample we would match (using **pidp**) to **a_indresp** and use the variable **a_jbstat**, while for the BHPS sample we would match (using **pid**) to **rindresp** and use the variable **rjbstat**.

4.5. USING THE "EXTRA 5 MINUTES" QUESTIONS

An "Extra 5 minutes" of question time is set aside for questions of particular interest to ethnicity related research (e.g., ethnic identity and remittances). To provide sufficient sample sizes to allow analysis of these questions separately by ethnic (minority) groups, the questions were asked of the EMBS (see Section 2.2.3), the GPC sample (see Section 2.2.2), and ethnic minority individuals living at Wave 1 in a Low Density Area (LDA). By LDA we mean areas where ethnic minority concentration was very low and which were not eligible for being selected for the EMB. In other words, members of ethnic minority groups living in these areas had a selection probability of 0 to be in the EMBS. The LDA ethnic minority status was fixed at Wave 1, i.e., changes in residence from Wave 2 onward do not affect membership in the "Extra 5 minutes" sample. From Wave 7 onward, IEMBS and non-UK born GPS sample members (status fixed by Wave 6) are also included in the "Extra 5 minutes" sample.

Those eligible for the "Extra 5 minutes" can be identified using the flag variable ***w*_xtra5min_dv**. Note that they also received the standard questionnaire that the rest of the sample received. The flag ***w*_xtra5minosm_dv** identifies OSMs who are eligible for these questions. In other words, ***w*_xtra5min_dv** is 1 and ***w*_xtra5minosm_dv** is 0 for TSMs who have joined the households of these OSMs who are eligible for the "Extra 5 minutes" questions.

To analyse this sample use the appropriate "Extra 5 minutes" (cross-sectional and longitudinal) weights (see Section 3.8). Note that there is no cross-sectional weight for the "Extra 5 minutes" questions as at Wave 2 these were only asked of sample members who had completed the main interview at Wave 1. Thus, the Wave 2 longitudinal weight should be used for Wave 2 cross-sectional analysis. "Extra 5 minutes" questions are listed under the Index Term "Extra 5 minutes questions" in the Online Data Documentation. Please also consult McFall, Nandi et al. (2018).

4.6. USING INFORMATION COLLECTED USING MIXED MODES

The mode used to administer a survey can affect the answers given by the survey respondents to the same questionnaire. Despite this possibility, the convenience and potential cost savings (especially relative to face-to-face/CAPI mode) have led Wave 8 to adopt a push-to-web mixed-mode design in which 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 (see Section 2.3.1.1 for further details). The implication of mode effects for Wave 8 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.

Before continuing, 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, socially 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 (see Section 2.3.1.1). 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.

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 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.

It is intended that future issues of the User Manual will be updated to reflect the results of the experimental analysis.

4.7. Example Code for Matching Files and Analysing Data

On our Study website we provide users with a number of examples of common data management and analysis tasks, see https://www.understandingsociety.ac.uk/documentation/mainstage/syntax.

We provide code in Stata and will be adding additional examples throughout the year. Some code is already translated into SPSS, and we will add to this as well as translating into SAS and R. We encourage users to get in touch with us via the user

support forum (see Section 6) if they have written code that they think others could benefit from or have suggestions for additional examples.

Our current code examples focus on:

- Distributing household level information to individuals
- Summarizing individual level information to household level
- Adding other household member's information to individual response data
- Using the egoalt data file to create household composition variables
- Merging individuals' responses from multiple waves into long format
- Merging individuals' responses from multiple waves into wide format
- Merging individual level files from the harmonised BHPS and UKHLS into long format
- Obtaining estimates that correctly take into account the sample design

Our online training course materials provide more detailed worked examples, see https://www.understandingsociety.ac.uk/help/training/online.

5. DATA ACCESS

The data are released through the UK Data Service (UKDS) in SPSS, Stata and CVS formats. While documentation is released through the UKDS, we encourage users to consult the *Understanding Society* webpage. The documentation will develop over time. We have developed specific guides about major content areas such as the biomeasures or cognitive measures, and guides for issues that are frequently problematic for users such as selection of appropriate weights. We will continue to develop specific user guides over time.

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 general release. Analysts may apply to gain access to restricted resources.

Users are required to sign licence agreements; the different agreements in place and which data they apply to, are described in the following sections.

Table 45 through Table 47 list the data products available under different licence agreements. The data can be accessed directly by replacing ## by the Study number in the following URL: https://discover.ukdataservice.ac.uk/catalogue/?sn=###.

5.1. RELEASE VERSIONS

5.1.1. END USER LICENCE (EUL) DATA

Most of the Wave 1 to 8 data has been released according to the conditions of the regular UKDS End User Licence (EUL): http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx#/tab-end-user-licence. The data are listed as SN 6614 - *Understanding Society*: Waves 1-8, 2009-2017.

Table 45: List of EUL data distributed through the UKDS

Study type[1]	Study no	Study title[2]
Core	6614	*Understanding Society*: Wave 1-8, 2009-2017
Link	7615	*Understanding Society*: Interviewer Survey, 2014

[1] "core" refers to data collected in the annual interviews; "link" refers to external data that have been added to the interview data using unique identifiers in the Study.

[2] Unless stated otherwise, the complete Study title begins with "*Understanding Society*: Wave 1-8, 2009-2017:"

5.1.2. SPECIAL LICENCE (SL) DATA

A number of sensitive data are released under Special Licence (SL). Researchers can apply for access to SL data through a UKDS application procedure, where they are required to justify their research objectives and explain why EUL data alone would be inadequate to reach those objectives; they are asked to report publications resulting from using the data. The conditions for using SL data are provided at http://ukdataservice.ac.uk/get-data/how-to-access/conditions/special-licence.aspx. Below, we briefly describe the different SL data products.

SN 6931- *Understanding Society*: Wave 1-5, 2009-2014: Special Licence Access is a copy of the EUL data (SN 6614) that contains the month of birth, full occupational coding, rare country of birth/nationality occurrences and uncapped income variables.

SN 7533 provides access to *Understanding Society* data linked with Department for Transport Accessibility Statistics 2009-2011 at the LSOA 2001 level. For further information on this file see Knies and Menon (2014). SN 7454, SN 7630, SN 6674, SN 7629, and SN 7453 each provide access to a neighbourhood classification which can be linked to *Understanding Society* data using the household identifier. Knies (2017) provides an overview of these classifications and how they may be exploited.

Some data products permit linking *Understanding Society* data via the wave-specific household identifier with published tables at the scale of official geographical units (SN 6666, SN 6668, SN 6671, SN 6675, SN 6672, SN 6673, SN 6669, SN 7182, SN 7245, SN 7249, SN 6670, SN 7248), or with information about schools (SN 7182). For further information about the geographical units included see ONS Geography website, http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/index.html.

Table 46: List of SL data distributed through the UKDS

Study type[1]	Study no	Study title[2]
Core	6931	*Understanding Society*: Wave 1-7, 2009-2016: Special Licence Access
Link	7533	*Understanding Society*: Waves 1-3, 2009-2012: Special Licence Access, Geographical Accessibility
Link	7454	Special Licence Access, Census 2001 Rural-Urban Indicators
Link	7630	Special Licence Access, Census 2011 Rural-Urban Indicators
Link	6674	Special Licence Access, Census 2001 Output Area Classification
Link	7629	Special Licence Access, Census 2011 Output Area Classification
Link	7453	Special Licence Access, Acorn Type 2015
link-id	6666	Special Licence Access, Local Authority District
link-id	6668	Special Licence Access, Westminster Parliamentary Constituencies
link-id	6671	Special Licence Access, Local Education Authorities
link-id	6675	Special Licence Access, Travel to Work Areas
link-id	6672	Special Licence Access, Strategic Health Authorities
link-id	6673	Special Licence Access, Primary Care Organisations
link-id	6669	Special Licence Access, Census Area Statistics Wards
link-id	7182	*Understanding Society*: Wave 1, 2009-2010: Special Licence Access, School Codes
link-id	7245	Special Licence Access, Census 2001 MSOA
link-id	7249	Special Licence Access, Census 2011 MSOA
link-id	6670	Special Licence Access, Census 2001 LSOA
link-id	7248	Special Licence Access, Census 2011 LSOA

[1] "core" refers to data collected in the annual interviews; "link" refers to external data that have been added to the interview data using unique identifiers in the Study. "link-id" refers to data files containing official identifiers that allow users to link their own data to the Study.
[2] Unless stated otherwise, the complete Study title begins with "*Understanding Society*: Wave 1-7, 2009-2016:"

5.1.3. ACCESS RESTRICTIONS TO SL DATA

Access to SL *Understanding Society* data products may be restricted, e.g., to UK-based users. The *Understanding Society* Data Access Strategy has further guidance on this: **https://www.understandingsociety.ac.uk/sites/default/files/downloads/general/2018/02/UKHLS_DAS_v23%20Nov2017.pdf.** Whether or not additional restrictions apply will be established in the application process, see Section 5.1.5 below. Data users start this process by applying to access the SL data that they require for their research project.

5.1.4. SECURE ACCESS

Some data can only be accessed in secure settings, which are supplied by the UKDS. Currently this covers two *Understanding Society* data products: SN 6676 includes postcode grid references and full date of birth and otherwise matches the SL data (SN 6931). SN 7642 contains information about children's education obtained linkage to official school records. For further information about the Secure Data Service, see http://ukdataservice.ac.uk/use-data/secure-lab.aspx.

Table 47: List of Secure Access data distributed through the UKDS

Study type[1]	Study no	Study title[2]
Core	6676	*Understanding Society*: Wave 1-8, 2009-2017 and Harmonised British Household Panel Survey (BHPS), Wave 1-18, 1991-2009: Secure Access
Link	7642	*Understanding Society*: Wave 1, 2009-2011: Linked National Pupil Database: Secure Access

[1] "core" refers to data collected in the annual interviews; "link" refers to external data that have been added to the interview data using unique identifiers in the Study. "link-id" refers to data files containing official identifiers that allow users to link their own data to the Study.

[2] Unless stated otherwise, the complete Study title begins with "*Understanding Society*: Wave 1-8, 2009-2017:"

5.1.5. TIMELINE FOR APPLICATIONS FOR SPECIAL LICENCE AND SECURE ACCESS

Applications to UKDS for Special Licences and Secure Data Access need to go through an additional review process by the *Understanding Society* team's Special Licence Officer, which is briefly outlined below, together with turnaround times.

1. Application made to access data via UKDS.
2. UKDS processes application and passes on to *Understanding Society* Special Licence Officer within three working days.
3. The *Understanding Society* Special Licence Officer then considers the application; if it is a particularly complex application it is reviewed by an internal access group.
4. Decisions are provided to UKDS within 10 working days; UKDS then responds to applicant within one working day.

The *Understanding Society* Governing Board reviews applications every six months; they also act as the appeal body for anyone is unhappy with their decision.

5.2. REVISIONS TO PREVIOUS RELEASES

We release the preceding waves of data when we make a new edition available. Users should refer to the document "UKHLS 2018 Revisions", which is supplied with the UKDS Study documentation
http://doc.ukdataservice.ac.uk/doc/6614/mrdoc/pdf/6614_ukhls_2018_revisions.pdf

We request that researchers using the data notify us about errors, inconsistencies, and other problems with the data identified during their use of the data. We make use of this information in improving the data.

Please raise any issues relating to data or data analysis with our Data User Support service at https://www.understandingsociety.ac.uk/support/projects/support.

We communicate information to members of the *Understanding Society* user group. Please register for the group using the registration form provided at the top right of the following page https://www.understandingsociety.ac.uk/support/projects/support.

The *Understanding Society* website has a Frequently Asked Questions (FAQ) https://www.understandingsociety.ac.uk/help/faqs. The Data User Support forum also has a FAQ about data-related questions: https://www.understandingsociety.ac.uk/support/projects/support/wiki.

5.3. LINKS TO OTHER STUDIES IN THE STUDY FAMILY

5.3.1. THE BRITISH HOUSEHOLD PANEL SURVEY

Understanding Society-harmonised BHPS data are included in the *Understanding Society* data release from the Wave 7 (November 2017) data release onward. Non-harmonised (stand-alone) data from the BHPS prior to joining *Understanding Society* can be obtained from the UK Data Service (SN5151 British Household Panel Survey, Waves 1-18, 1991-2009: https://discover.ukdataservice.ac.uk/catalogue?sn=5151. The Study documentation is available at http://www.iser.essex.ac.uk/bhps.

For users interested in analysing data from both *Understanding Society* and the BHPS we recommend reading the *Understanding Society* harmonised British Household Panel Survey User Guide (Fumagalli, Knies et al. 2017), Section 1, to inform their choices over using the stand-alone or harmonised BHPS data.

5.3.2. THE WAVES 2-3 NURSE HEALTH ASSESSMENT

In 2010-2012, *Understanding Society* augmented survey questions with direct health assessments and the collection of blood samples, see McFall, Petersen et al. (2014). The blood samples were subsequently analysed to produce a range of biomarkers, see Benzeval, Davillas et al. (2014). Data from the Wave 2 and Wave 3 health assessment, including the blood-based biomarkers, are released through the UKDS SN7251 *Understanding Society*: Waves 2-3 Nurse Health Assessment, 2010-2012: http://discover.ukdataservice.ac.uk/catalogue?sn=7251. Note that SN 7587 contains more sensitive information (such as detailed socio-economic classifications) and is available under SL.

5.3.3. GENETICS DATA

With consent, DNA was extracted from the blood samples taken at the nurse visit and analysed using the Illumina Human Core and Exome chip. These data can be accessed in two ways.

Genetics-only data are available directly from the European Genome-phenome Archive. Researchers who ONLY wish to access the genome wide scan data should apply to do so directly at the European Genome-phenome Archive, see https://www.ebi.ac.uk/ega/studies/EGAS00001001232, and new applications will be considered by the Wellcome Trust Sanger Institute's Data Access Committee.

Researchers who wish to access genetics data combined with *Understanding Society* survey data need to apply to the ESRC's METADAC. Details of the

application process can be found on the METADAC website, see http://www.metadac.ac.uk/data-access-committee/.

Further information about both of these application processes is available on the *Understanding Society* website.

5.3.4. THE UNDERSTANDING SOCIETY INTERVIEWER SURVEY 2014

The *Understanding Society*: Interviewer Survey 2014 provides information from interviewers who had worked on *Understanding Society* at Wave 1 (which took place in 2009-2010), including those interviewers who had since left the employ of NatCen. Whilst the standard version of *Understanding Society* contains basic objective demographic information on interviewers (sex, ethnicity, years of experience at NatCen, age), the survey focused on subjective measures; attitudes and opinions.

The data has been released through the UKDS SN7615 *Understanding Society:* Interviewer Survey 2014, SN7615: http://discover.ukdataservice.ac.uk/Catalogue/?sn=7615.

5.3.5. THE UNDERSTANDING SOCIETY INNOVATION PANEL

The *Understanding Society* project incorporates the Innovation Panel (IP), a separate survey intended to support methodological research; see https://www.understandingsociety.ac.uk/documentation/innovation-panel. Data from the IP has been released through the UKDS SN6849 *Understanding Society*: Innovation Panel, Waves 1-10, 2008-2017: http://discover.ukdataservice.ac.uk/Catalogue/?sn=6849.

5.3.6. THE CROSS-NATIONAL EQUIVALENT FILE (CNEF)

Understanding Society is part of a world-wide family of household panel studies and included in the Cross National Equivalent File (CNEF), the result of a cooperative effort of individuals and organisations involved in the collection of household panel data. The CNEF data contains information common to surveys from at least two of the following eight countries: the Household Income and Labour Dynamics in Australia (HILDA) survey for Australia, the Survey of Labour and Income Dynamics (SLID) for Canada, the Socio-Economic Panel Study (SOEP) for Germany, the Korea Labor and Income Panel Study (KLIPS) for South Korea, the Russia Longitudinal Monitoring Survey (RLMS-HSE) for Russia, the Swiss Household Panel (SHP) for Switzerland, the BHPS and *Understanding Society* for the United Kingdom, and the Panel Study of Income Dynamics (PSID) for the United States of America. Detailed information about the project, including codebooks, can be found on the designated project website https://cnef.ehe.osu.edu/data/.

Data from the CNEF can be accessed by filling an application form (requirements for data access change by country). Upon approval of the application, the data for which the access has been granted are sent to users in a password protected file. More information can be found here: https://cnef.ehe.osu.edu/data/access-procedures/.

5.4. 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. Table 48 provides information on the various committees which have provided ethical approval of the *Understanding Society* study and its components as appropriate.

Table 48: Information on ethical reviews of the Study and its components

Main survey
Ethics Committee of the University of Essex: 6 July 2007 (Waves 1 to 2) 17 December 2010 (Waves 3 to 5) 20 August 2013, 31 July 2014, 1 July 2015, and 29 February 2016 (Waves 6 to 8)
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 anuary 2011

6. ONLINE DATA USER SUPPORT AND RESOURCES

Understanding Society has a wealth of information online at:
https://www.understandingsociety.ac.uk/

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.

Further Help and Support for using *Understanding Society* can be found in the Online Data User forum. After a short registration data users can read past issues, FAQ's and experiences or report any issues or queries of their own.

The URL is: https://www.understandingsociety.ac.uk/support/projects/support

Users should read the "How to raise an issue" guidance before posting a question. We aim to respond to all queries within 10 working days.

Users may also email user support directly using our email:
usersupport@understandingsociety.ac.uk

Our preferred mode of communication is via the forum as other users may then also benefit from the information provided.

The team at *Understanding Society* are also offering online training courses which provide plenty of worked examples of how to prepare and analyse *Understanding Society* data. Presently the following courses are available on Moodle:

- Introduction to *Understanding Society* Using Stata (also: using SPSS or SAS)

- *Understanding Society* for Transport Analysis (using Stata),
- Introduction to British Household Panel Studies (BHPS) using Stata

For an up-to-date list see,

https://www.understandingsociety.ac.uk/help/training.

In person training courses are also available both for general introductions to the Study and specialised aspects, such as weighting, biomarkers, IP, genetics. Further information can be found on the website.

7. CITATIONS AND ACKNOWLEDGEMENTS

Any publication, whether printed, electronic or broadcast, based wholly or in part on the *Understanding Society* data collection provided by the UK Data Service must be accompanied by the correct citation and acknowledge the Institute for Social and Economic Research as the data provider and the UK Data Service as the data distributor. The acknowledgement, which gives credit to sponsors or distributors, is not a replacement for a proper citation. We recommend the following wording:

> "*Understanding Society* is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service."

7.1. CITATION OF THE DATA

The format for bibliographic references is as follows:

> University of Essex. Institute for Social and Economic Research, NatCen Social Research, Kantar Public (2018): *Understanding Society: Waves 1-8, 2009-2017 and Harmonised BHPS: Waves 1-18, 1991-2009.* [data collection]. 11th Edition. UK Data Service. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-12.

7.2. CITATION OF THE USER GUIDE

The User Guide is to be cited as follows:

> Knies, Gundi (ed.) (2018). *Understanding Society*: Waves 1-8, 2009-2017 and Harmonised BHPS: Waves 1-18, 1991-2009, User Guide, November 2018, Colchester: University of Essex.

7.3. ACKNOWLEDGMENTS

People who participated in writing sections of the documentation for this or prior releases include, in alphabetical order, Gina Anghelescu, Stephanie Auty, Randy Banks, Yanchun Bao, Michaela Benzeval, Nick Buck, Jon Burton, Paul Clarke, Paul Fisher, Laura Fumagalli, Olena Kaminska, Gundi Knies, Peter Lynn, Stephanie McFall, Alita Nandi, Victoria L. Nolan and Jakob Petersen. A very big "Thank you!" goes to the many people have and continue to contribute to the unrelenting success and timely delivery of *Understanding Society*. They not only include the Study directors and scientific advisors (see

https://www.understandingsociety.ac.uk/about/team), but also the many interviewers who collect the data on behalf of ISER and the fieldwork agencies, the Study members who provide their information, the people who maintain contact with Study members, prepare the data for release, provide user support, organise training events and inform the public about new findings from the Study.

8. REFERENCES

Al Baghal, T., A. Jaeckle, et al. (2015). Understanding Society -The UK Household Longitudinal Study: Innovation Panel, Waves 1-8, User Manual. Colchester, ISER University of Essex.

Benzeval, M., A. Davillas, et al. (2014). Understanding Society:The UK Household Longitudinal Study: Biomarker User Guide and Glossary. Colchester, University of Essex.

Berthoud, R., L. Fumagalli, et al. (2009). "Design of the Understanding Society Ethnic Minority Boost Sample." Understanding Society Working Paper **2009-02**.

Budd, S., E. Gilbert, et al. (2012). Understanding Society Innovation Panel Wave 4: Results from Methodological Experiments. Understanding Society Working Paper. J. Burton. Colchester, Institute for Social and Economic Research. **2012-06**.

Carpenter, H. and J. Burton (2018). "Adaptive push-to-web: experiments in a household panel study." Understanding Society Working Paper Series(2018-05).

D'Ardenne, J., D. Collins, et al. (2017). "Assessing the risk of mode effects: review of proposed survey questions for waves 7-10 of Understanding Society." Understanding Society Working Paper **2017-04**

Fumagalli, L., G. Knies, et al. (2017). Understanding Society: The UK Household Longitudinal Study harmonised British Household Panel Survey (BHPS) User Guide. Colchester, University of Essex. Institute for Social and Economic Research.

Hayes, C. and H. Watson (2009). "HILDA Imputation Methods." HILDA Project Technical Paper Series, University of Melbourne **No 02/9**.

Kenward, M. and J. Carpenter (2007). "Multiple imputation: current perspectives." Statistical Methods in Medical Research **16**(3): 199-218.

Knies, G. (2017). "Exploring the Value of Understanding Society for Neighbourhood Effects Analyses." Research Data Journal for the Humanities and Social Sciences.

Knies, G. and S. Menon (2014). "Understanding Society: Waves 1-3, 2009-2012: Special Licence Access, Geographical Accessibility. The UKHLS-Accessibility Data File User Guide " http://doc.ukdataservice.ac.uk/doc/7533/mrdoc/pdf/7533_ukhls_accessibility_userguide.pdf.

Little, R. J. A. (1988). "Missing Data Adjustments in Large Surveys." Journal of Business and Economic Statistics **6**(287-296).

Little, R. J. A. and H. L. Su (1989). Item Non-response in Panel Surveys. Panel Surveys. D. Kasprzyk, G. J. Duncan, G. Kalton and M. P. Singh. New York, Wiley.

Lynn, P. (2009). "Sample design for Understanding Society." Understanding Society Working Paper **2009-01**.

Lynn, P., J. Burton, et al. (2012). "An initial look at non-response and attrition." Understanding Society Working Paper **2012-02**.

Lynn, P., A. Nandi, et al. (2016). Design and implementation of a high quality probability sample of immigrants and ethnic minorities. Colchester, University of Essex.

Lynn, P. L. and G. Knies (2015). Understanding Society Quality Profile. Understanding Society Quality Profile. P. L. Lynn and G. Knies. Colchester, ISER University of Essex. **1**.

McFall, S. (2013). "Understanding Society- UK Household Longitudinal Study: Cognitive Ability Measures." Understanding Society User Manual.

McFall, S., A. Nandi, et al. (2018). Understanding Society: UK Household Longitudinal Study: User Guide to Ethnicity and Immigration Research. Colchester, ISER University of Essex.

McFall, S., J. Petersen, et al. (2014). "Understanding Society Waves 2 and 3 Nurse Health Assessment, 2010-2012: Guide to Nurse Health Assessment." https://www.understandingsociety.ac.uk/documentation/health-assessment.

Office for National Statistics (1991). Census: Key Statistics for Local Authorities. London, HMSO.

Office for National Statistics (2003). "Census 2001: Key Statistics for Local Authorities in England and Wales'." www.ons.gov.uk/ons/rel/census/census-2001-key-statistics/local-authorities-in-england-and-wales/.

Ragunathan, E. T., J. M. Lepkowski, et al. (2001). "A Multivariate technique for multiply imputing missing values using a sequence of regression models." Survey Methodology **27**(1): 85-95.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York, Wiley.

Schafer, J. (1997). Analysis of Incomplete Multivariate Data. London, Chapman & Hall.

Spanier, G. B. (1976). "Measuring dyadic adjustment: new scales for assessing the quality of marriage and similar dyads." Journal of Marriage and the Family **38**: 15-27.

Taylor, M. F. (2010). British Household Panel Survey User Manual Volume A: Introduction, technical report and appendices. Colchester, University of Essex.

Taylor, M. F. (2010). Weighting, imputation and sampling errors. British Household Panel Survey User Manual Volume A: Introduction, technical report and appendices. J. Brice, N. Buck and E. Prentice-Lane. Colchester, University of Essex**:** A5 1-13.

van Buuren, S., H. C. Boshuizen, et al. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis." Statistics in Medicine **18**: 681-694.