Understanding Society and its income data

Paul Fisher
Laura Fumagalli
Nick Buck
Silvia Avram

University of Essex

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Non-technical summary

Understanding Society - the UK Household Longitudinal Study (UKHLS) follows respondents through time and collects detailed information on their income at each point in time. Derived from these survey reports of income, and included in the publically available UKHLS datasets, are a number of variables corresponding to different income concepts. The derived variables correspond to individual level income variables, household level income variables, and household level deductions (council tax and housing costs). It is important for users to know not only what income variables are available in UKHLS, but also to know how they were calculated. More generally, users of UKHLS income data also need some background to the study. The aim of this paper is to provide such information to fill this gap.

The paper describes the key features of UKHLS including: its large sample size and subsamples, the timing of data collection, response rates including responses rates to income questions, attrition, statistical weighting, sample definitions and following rules.

On the UKHLS income variables, the paper records a whole range of information that is required by users. It describes the questionnaire content on income and its evolution over time; discusses equivalisation and price indices; summarises UKHLS differences with its predecessor the British Household Panel Survey; and describes the possibilities for administrative income data linkage. The paper also carefully describes how derived income variables are constructed, for example, how missing data are treated, and the simulation of after tax incomes when only gross incomes are observed.

UKHLS is a critical data source as it is one of the few UK data sets that contains information on individuals' income and also follows these individuals over time. For example, UKHLS is the data source for the Department for Work and Pensions publication on income dynamics. It is the source for academic publications on the income distribution. Many other research uses of UKHLS do not have income as their central focus, but nonetheless, analyse UKHLS income variables. For these reasons, it is highly desirable to monitor the quality of UKHLS income data and how it changes over time. Therefore, as a quality assessment, the paper also compares UKHLS estimates of net household incomes to those from a reliable cross-sectional income source (the Households Below Average Income Series (HBAI)). The two sets of estimates are reassuringly similar to each other. We therefore conclude that UKHLS income data is of high quality and an excellent data source for research on the income distribution.

Understanding Society and its income data

Paul Fisher, Laura Fumagalli, Nick Buck and Silvia Avram

University of Essex

Abstract

The first aim of this paper is to describe information about Understanding Society - the UK Household Longitudinal Study (UKHLS) - that is relevant for research using its income data. UKHLS contains a wide range of income variables (individual level income variables, household level income variables, and household level deductions) and this paper documents the details of all of them. The second aim is to establish the quality of the UKHLS net household income data, relative to a cross-sectional benchmark. We find that estimates from the UKHLS line-up closely to those from the data source for UK official income statistics (Households Below Average Income).

Keywords: Understanding Society, income distribution, inequality, poverty
JEL Classification numbers: C83, D31, I32

Corresponding author: Paul Fisher, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom. Email: pfishe@essex.ac.uk

Contents

1. INTRODUCTION

Our aims in this paper are twofold. First, we wish to document information about Understanding Society - the UK Household Longitudinal Study (UKHLS) - that is relevant for research using its income data.[1] We hope that the information we document will act as a useful resource for users of the data. Our second aim is to establish the quality of the UKHLS net income data, relative to a cross-sectional benchmark. The latter is of much importance as UKHLS is one of the few UK sources of longitudinal income data and so it is critical to establish its quality. For example, UKHLS is the source for UK government statistics on the dynamics of low income (DWP 2018) and it has recently been used in academic research to document trends in income volatility (Avram, et al. 2019). Separately, it is the data source for a vast number of research papers where income plays a central role, even if it is not the main outcome variable of the analysis. Put together, we believe the importance of establishing the quality of UKHLS income data is clear. We organise the paper into 5 broad sections.

Section 2 gives a brief overview of the overall UKHLS study design and implementation. It describes the key features of the study relevant for research on income including: its large sample size and multiple subsamples, the timing of data collection, response rates including responses rates to income questions, attrition, weighting, sample definitions and following rules.

Section 3 discusses in detail issues directly related to the collection and analysis of income data in UKHLS: It describes the questionnaire content on income and its evolution over time; carefully records the construction of the derived net income variables; discusses the use of equivalisation and price indices; summarises UKHLS differences with the former British Household Panel Survey; and describes the potential for UKHLS to be linked to administrative income records in the future.

Section 4 sets out the UKHLS approach to dealing with missing data. Missing data are addressed by imputation and the section sets out the UKHLS imputation methods in detail, and shows the effects of imputation on the final income distributions.

Section 5 summarises the final UKHLS income derived variables. Several income derived variables are publically available and so, this is a good place for new users of the data to start. Section 5 organises the variables according to: individual level income variables, household level income

---

[1] 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. The version of the data used in this paper is: University of Essex, Institute for Social and Economic Research. (2018). *Understanding Society: Waves 1-8, 2009-2017 and Harmonised BHPS: Waves 1-18, 1991-2009*. *11th Edition.* UK Data Service. SN: 6614, http://doi.org/10.5255/UKDA-SN-6614-12

variables, and household level deductions (council tax and housing costs). The section provides variable names and a summary table for ease of reference to the data.

Finally, section 6 assesses the quality and representativeness of the net household income data. UKHLS estimates of the net income distribution are compared to those from a cross-sectional specialist income survey (the Households Below Average Income Series (HBAI)). The two sets of estimates line-up remarkably closely. We see similar levels and trends for various income percentiles, inequality measures, the composition of household incomes across the distribution, and the types of benefit unit we see across the distribution. Where there are differences, we see that UKHLS, relative to HBAI, tends to slightly overestimate income at the bottom of the distribution and underestimate it slightly at the top. This means that overall UKHLS tends to estimate slightly lower inequality levels. We conclude that UKHLS is a sound basis for both cross-sectional and longitudinal research on the income distribution.

2. UKHLS STUDY DESIGN

This section describes the key features of the study. It covers: the sample design; household and family definitions used; the following and eligibility rules; the timing and mode of data collection; and finally details of non-response (to income questions and overall response rates) and weighting.

2.1. OVERVIEW

The UKHLS built on the success of the British Household Panel Survey (BHPS).[2] The BHPS has been heavily used by government departments and by researchers within and outside the UK. It has been accessed by more than 5000 users and generates more than 150 publications per year. The ambition of the UKHLS has been to support a much wider range of research than the BHPS, based on a much larger sample size and a wider range of data collection. The collection of UKHLS started in 2009. Sample members are interviewed annually. As of July 2019, eight waves of data are currently available to researchers, with an ninth to be released this autumn. The study incorporates the BHPS sample, for which 25 waves of data are now available (Fumagalli, Knies and Buck 2017).

The UKHLS forms part of an international network of longitudinal studies including the Panel Study of Income Dynamics (Hill 1992), the German Socio-Economic Panel Study (Wagner, Burkhauser and Bheringer 1993), the Household, Income, and Labour Dynamics in Australia Survey (Wooden, Freidin and Watson 2002), the Swiss Household Panel (Budowski, et al. 2001), the Survey of Labour and Income Dynamics in Canada (Webber 1994) and other active household panels in South Africa, Israel, Korea, and China. The household panel design was established in the Panel Study of Income Dynamics (PSID) in the USA in the late 1960s. This design has proved extremely powerful in understanding the dynamics of populations and the determinants of behaviour and outcomes at household and individual level.

The design of UKHLS follows the standard household panel design. It involves selecting all individuals found in an initial representative sample of households (Original Sample Members or OSM) and following those individuals and their household members over time. The design provides repeated measures (longitudinal data) about OSMs and their children, but also data from other members of households in which OSM reside over time, allowing high quality household context measures to be used in analysis. A key feature of the household panel design is that the initial sample is representative of the whole population of all ages, and, with appropriate following rules, it will remain representative of the population as it evolves over time, except for new immigrants to the

---

[2] This paper is modelled loosely on Jenkins (2011), which discusses BHPS income data.

population. Research from studies with a household panel design supports direct inferences about the whole population.

The longitudinal elements are the individual people. The UKHLS is not a longitudinal study of households, since arguably households have no coherent existence over time, and focusing analysis only on households whose composition does not change between waves leads to severe biases (Duncan and Hill 1985). Rather, the UKHLS is a study of individuals in their changing household contexts and this context is very important for analysis of many life domains (Giles 2001).

The design of UKHLS is particularly relevant for the analysis of income, and economic well-being more generally, since it allows the collection of high-quality longitudinal measures of household income. This permits to analyse the transitions in economic circumstances experienced by the members of the population, such as the persistence of poverty or the factors affecting income mobility. For example, data from BHPS and UKHLS based on this design are regularly used in UK government publications on the dynamics of low income (DWP 2018).

In addition to the main survey, there is a separate survey, the Innovation Panel (IP), which is fielded in the year before the main survey. It is used to test innovative ways of collecting data and develop new areas of research. It contains the same main measures as the main survey, although its instruments differ somewhat. The IP has been used for experimental work on improving the measurement of income and related measures. Details can be found in Benzeval et al. (2017).

2.2. SAMPLE DESIGN

The *Understanding Society* main survey sample consists of a 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). Details of the sample design of the GPS, EMBS, and BHPS components can be found in Lynn (2009). Details of the sample design of the IEMB sample can be found in Lynn, Nandi et al. (2018).

GENERAL POPULATION SAMPLE

The general population sample (GPS) is based upon two separate samples of residential addresses in England, Scotland and Wales (Great Britain), and in Northern Ireland. The Great Britain sample is a proportionately stratified (equal probability), clustered sample of addresses selected from the

Postcode Address File.[3] Within Great Britain, the Primary Sampling Units (PSUs) are postal sectors (stratified by nine regions of England, plus Scotland and Wales), population density and minority ethnic density. 2,640 postal sectors were selected systematically, with probability proportional to size (number of addresses). Within each sampled sector, 18 addresses were selected systematically, resulting in an equal-probability sample of a total of 47,520 addresses in Great Britain. The Northern Ireland sample is not clustered. In Northern Ireland, 2,400 addresses were selected systematically from the Land and Property Services Agency list of domestic properties, thus making a total of 49,920 selected addresses in the UK. The NI sample had twice the selection probability of the GB part of the GPS.

Since constraints of survey capacity meant that fieldwork needed to be spread over a two-year period, the overall sample was divided into 24 monthly sub-samples, each independently representative of the UK population. This means that differences over time within a wave can be compared using nationally representative samples, and annual or quarterly subsets can be independently analysed.

FORMER BHPS SAMPLE

Understanding Society incorporates the BHPS sample members into the overall sample design beginning in Wave 2. From wave 2 onwards, the BHPS sample has the same questionnaire as the Understanding Society general population sample. The extensive longitudinal data of the BHPS has great scientific value, including the opportunity for early longitudinal analyses of *Understanding Society*. The BHPS was a random sample of Great Britain, excluding the Scottish Highlands and Islands. In its first wave in 1991, it achieved a sample of 5,500 households. Boost samples of Scotland and Wales were added in 1999 and of Northern Ireland in 2001. These modifications were motivated by interest in analyses in these countries, related to political changes associated with devolution in the UK. For an excellent discussion of the former BHPS and its income data see Jenkins (2011).

ETHNIC MINORITY BOOST SAMPLE

The goal for the ethnic minority boost sample was to provide samples of at least 1,000 adults in each of the five largest ethnic minority groups: Indian, Pakistani, Bangladeshi, Caribbean and African. Such

---

[3] Postcodes are codes referring to a group of UK postal address. All UK postcodes are stored in a national database known as the Postcode Address File.

a sample supports group-specific analyses of these ethnic groups (Berthoud, et al. 2009). While the sampling targets are defined in terms of numbers of adults, the sample is of households.

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. These high-density 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. Further sub-sampling of the high-density areas was done to increase the efficiency of the yield: a higher sampling fraction was used for areas expected to yield three or more households, while successively smaller fractions were used for areas expected to yield two, one or zero ethnic minority households.

At selected addresses, households were screened for the presence of a member of a minority ethnic group. The screening question was, "Do you come from or have parents or grandparents who come from any of the following ethnic groups?" The response categories were: Indian, mixed Indian, Pakistan, Bangladeshi, Sri Lankan, Caribbean/West Indian, mixed Caribbean/West Indian, North African, Black African, African Asian, Chinese, Far Eastern, Turkish, or Middle Eastern/Iranian, or other. At the screening stage, all households with the smaller ethnic groups were selected and there is some de-selection of larger ethnic minority groups, e.g. Indians. White minorities were not selected in the screening for the EMBS, but can instead be found in the GPS. Following the first six months of data collection the procedures were reviewed and modified. The more significant change was increasing the number of addresses issued in areas estimated to be high in Bangladeshi: the smallest of the five main ethnic groups. This was because the number of interviews with Bangladeshis was lower than expected and it was at risk of failing to meet the 1000 target (Berthoud, et al. 2009).

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.

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 sampled from the Postcode Address File.[2] The sample follows 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). At the doorstep the interviewer asked "Is there anyone living at this address who was born outside the United Kingdom, including children?" If the answer was yes, the household was screened in. If the answer was no, they were asked the EMBS screening question (see above). Thus 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 EMB and the IEMB do not 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.

2.3. HOUSEHOLD AND FAMILY DEFINITIONS

The UKHLS definition of a household follows UK statistical practice and is the same as that used in UK government surveys. A household is 'one person living alone or a group of people who either share living accommodation or share one meal a day and who have the address as their only or main residence'. The definition also requires six months' continuous residence, implying that students will be included at their term time address, unless living at a hall of residence.

UKHLS does not work with a family definition in the way that PSID does, for example, as its primary unit of analysis. However, UKHLS data permits to identify units below the household level consistent with those used in the British tax and benefit system and known as benefit units. These are defined as a single person or a couple living together and would include dependent children of adult

members of the benefit unit.[4] These units are used for the assessment of means tested state benefits. A household may combine several benefit units, for example where a non-dependent child lives with his or her parents. Figure 1 shows the relationship between individuals, benefit units and households graphically

Figure 1: Individuals, benefit units and households
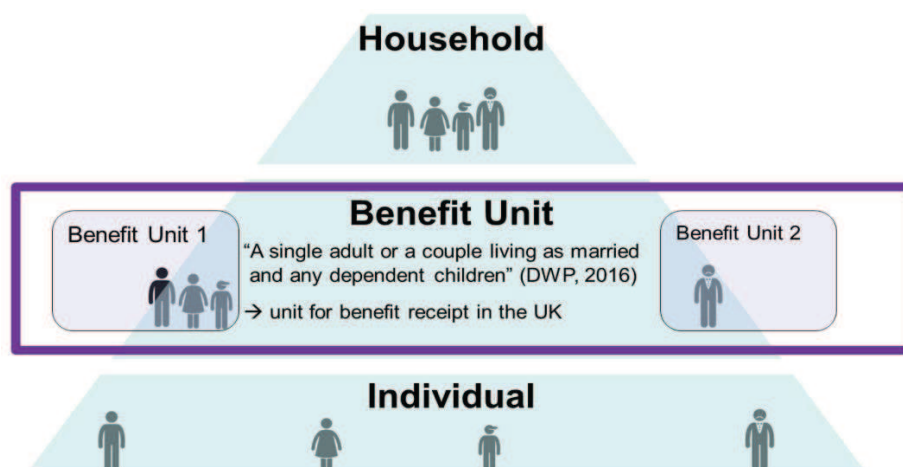
2.4.   FOLLOWING RESPONDENTS OVER TIME AND ELIGIBILITY FOR INTERVIEW

The rules used to assign sampling statuses, and thus decide who is included in the sample after wave one and then followed over time, are called following rules. The following rules mimic the demographic processes by which the population is reproduced, including births and deaths, partnership formations and dissolutions, and emigration. They provide a natural sampling method over time, which represents the evolving pattern of households and families in the UK. The one exception is that there is no direct way in which the following rules capture immigrants into the UK. Apart from immigration, the sample remains representative of the UK population as it changes over time, subject to adjusting for attrition. Whether and how to sample new immigrants remains an issue to be decided in the future development of the study.

Notionally, the individuals found at selected households in the first wave are designated as Original Sample Members (OSM) and we attempt to retain OSM respondents as part of the sample as long as they live in the UK. Individuals joining the household of an OSM after the sample selection/first interview are temporary sample members (TSM). However, births to an OSM Mother are also classified as OSMs. We attempt to interview TSM participants in successive waves as long as they

---

[4] Dependent children are aged under 16 or under 19 and in full-time education, excluding higher education, and not married or with a child.

live in the household of an OSM. In sum, TSMs are not followed for interviews when they leave the household, but OSMs are. We provide precise details of sample statuses in the additional materials (section 7.1).

In general, longitudinal analysis of individuals will focus on OSM respondents only, since TSMs drop out when they no longer live with OSMs. However, data from TSM respondents are used to compute household measures, such as household income. Typically, analysis of income mobility uses both OSM and TSM data to construct household income measures. However, the analysis is only carried out on OSM and household income is treated as an attribute of these individuals.

2.5. TIMING AND MODE OF DATA COLLECTION

As a result of fieldwork capacity issues given the very large sample size, particularly in the early years, data collection for each wave of UKHLS covers a 24-month period, and individual waves overlap so that sample members are interviewed at annual intervals. The sample is issued in 24 monthly samples and sample members remain in the same month of issue at each wave, so that the interval between interviews should be approximately 12 months. The field period for each month sample lasts for around four months to track movers and undertake refusal conversion, so there may be some variation in the interval between individual interviews.

There are some specific exceptions to the 24-month fieldwork period. In planning the timing of fieldwork for the BHPS sample component, it was necessary to balance fully integrating the sample into Understanding Society as against creating a discontinuity in the BHPS series. After consultation, it was decided that it was most important to ensure the integration of BHPS into the new study (Laurie 2010). Therefore, the fieldwork for the BHPS sample component was distributed evenly over the 12 months of the first year of data collection beginning in January 2010, as part of wave 2 of Understanding Society.[5] This introduces a one-off longer gap between interviews for the BHPS sample. The Northern Ireland sample is also issued over the first 12 months of a wave. Table 1 shows the timing of data collection for waves one to eight.

Most of the data collection is conducted face-to-face via computer assisted personal interview (CAPI). There are also self-completion instruments for youth and adults. The youth self-completion questionnaire is administered on paper. The adult self-completion questionnaire shifted from paper to computer assisted self-interview (CASI) in Wave 3. From Wave 3 onwards, there was also a telephone mop-up at the end of the fieldwork period for each sample month. At Wave 7, for the first

---

[5] Up to 2008, the fieldwork for BHPS was concentrated between September and December.

Table 1: Timing of data collection waves 1-8

Release markers: w1 release, w2 release, w3 release, w4 release, w5 release, w6 release, w7 release, w8 release

Year: 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018

GPS GB + EMB
- wave one
- wave two
- wave three
- wave four
- wave five
- wave six
- wave seven
- wave eight

GPS NI
- wave one
- wave two
- wave three
- wave four
- wave five
- wave six
- wave seven
- wave eight

BHPS
- wave two
- wave three
- wave four
- wave five
- wave six
- wave seven
- wave eight

IEMB
- wave six
- wave seven
- wave eight

time, households where no member has participated at wave 6 were offered the chance of completing the interview by computer assisted web interviews (CAWI). The web interview was conducted using the same questionnaire for the CAPI interviews, slightly modified to accommodate the fact that an interviewer was not present. From wave 8, a larger proportion of respondents (40% at wave 8) are offered an interview by web as a core data collection mode.

The interested reader can find further details on the overall study design in a number of other sources, including the UKHLS User Manual (Knies 2017) and the UKHLS Quality Profile (Lynn and Knies 2015).

2.6. RESPONSE OUTCOMES

Response outcomes can be considered in four different ways. Firstly, there is the question of whether a household containing sample members provides any response. This is consistent with the way in which cross-sectional household surveys report response. Secondly, and most important for longitudinal analysis, is the question of whether individual sample members continue to respond at successive waves or drop out, so-called attrition. Thirdly, there is the question of whether all eligible members of participating households provide a response. This is important for the construction of household level measures, including household income. Finally, there is the issue of response to individual questions for sample members who do provide some response to the questionnaire, so-called item response or non-response. This section briefly discusses the last three of these types of response. The first, household level response, is discussed in Lynn and Knies (2015). Lynn and Borkowska (2018) also provide some analysis of overall sample representativeness.

ATTRITION

For each sample component, table 2 shows wave-on-wave retention rates for full adult respondents, that is the percentage of those interviewed in a given wave who are also interviewed the next wave. Three aspects should be noted. First, the retention rates are clearly lower than in the early waves of BHPS, reflecting the much greater difficulty of achieving very high response rates in recent years compared with the early 1990s. A similar decline is reflected in other major UK and international surveys (Meyer, et al. 2015). Second, retention rates are lower in the EMBS and the IEMB. Third, there is a drop-in response in wave 6, associated with a change in fieldwork organisation (see section 3.3).

Table 2: Wave-on-wave retention rates: adult respondents

	General Population Sample	BHPS	Ethnic Minority Boost	Immigrant and Ethnic Minority Boost	All
Wave 1 to wave 2	76.22	-	62.42	-	74.29
Wave 2 to wave 3	80.03	85.32	71.25	-	80.38
Wave 3 to wave 4	84.70	85.33	76.17	-	84.03
Wave 4 to wave 5	86.94	87.82	79.20	-	86.40
Wave 5 to wave 6	82.60	85.09	74.15	-	82.38
Wave 6 to wave 7	86.05	87.09	78.97	62.87	83.24
Wave 7 to wave 8	86.54	88.81	80.90	69.51	85.12

Table 3 shows retention rates for all enumerated individuals and so indicates whether an individual in a productive household is productive again at the next wave. It should be noted that, for both respondents and enumerated individuals, there are significant numbers who drop out for one wave but are re-interviewed at the subsequent wave. Therefore, the cumulative pattern of attrition is not simply the product of the individual wave attrition rates.

Table 3: Wave-on-wave retention rates: enumerated individuals

	General Population Sample	BHPS	Ethnic Minority Boost	Immigrant and Ethnic Minority Boost	All
Wave 1 to wave 2	77.25	-	66.45	-	75.38
Wave 2 to wave 3	82.28	87.30	77.63	-	82.79
Wave 3 to wave 4	85.89	86.79	80.30	-	85.40
Wave 4 to wave 5	87.89	88.81	82.89	-	87.49
Wave 5 to wave 6	84.38	87.01	80.94	-	84.56
Wave 6 to wave 7	86.99	88.09	82.43	66.11	84.04
Wave 7 to wave 8	87.86	89.70	83.55	72.13	86.22

Ideally, we want all adult members of participating households to respond to the full questionnaire. For income analysis, this is desirable to be able to compute total household income from reports of all members. Unfortunately, it is difficult to obtain complete response for all households, because some members are often unable or unwilling to respond. Where the household member has not refused to be involved in the study, proxy data may be collected. This contains some limited information about personal income. Where the member has refused to be involved or no-one can provide proxy information, information is restricted to that collected on the household membership roster.

Table 4 shows the extent of within household non-response in UKHLS at wave 8, the latest wave at the time of writing. Differences across waves are relatively small, and thus are not presented. There is some difference between sample components, with lower complete household response in the EMBS and the IEMBS. The BHPS has a somewhat higher complete household response, though this is not as high as it was when the BHPS was being conducted on its own.[6]

Table 4: Percentage distribution of households by within household response, Wave 8

	General Population Sample	BHPS	Ethnic Minority Boost	Immigrant and Ethnic Minority Boost	All
Complete response: all eligible adults interviewed	78.35	81.29	68.86	72.74	72.74
All eligible adults interview or proxy	5.84	5.10	11.00	7.60	7.60
At least one within household refusal	15.81	13.61	20.14	19.65	19.65

There are two approaches to the analysis of household income in the presence of incomplete household response. The first is to restrict the analysis to complete households only, and possibly reweight for the biases which may result from the selection. The second is to impute incomes for the proxy and individual non-respondents in responding households and create a total household income based on these imputations. UKHLS data production processes support the second

---

[6] For example, at wave one of the BHPS 87 percent of households had complete response and 94 percent had either complete response or complete response including proxy interviews.

approach, as discussed in the section on imputation. As indicated above, the incomplete households have different characteristics from those in which all members respond, and it is recommend that if the analysis of household income is restricted to complete households then results should be reweighted. A specific weight for this purpose is not presently provided in the public release data.

ITEM NON-RESPONSE

Average item non-response rates are relatively low, with around 1% of items missing on average in the individual interview administered schedule. However, income questions are often seen by respondents as sensitive and may be asking for information respondents do not necessarily know without referring to documents. As a result, item non-response rates on income variables tend to be substantially higher than the average level. This is shown in figure 2 below which presents item non-response rates for four key income variables. There is some tendency for item non-response rates to drop as participants continue in the survey for longer, although a slight increase in item non-response was seen from wave 6, when the fieldwork agency changed. Rates are broadly comparable with those experienced in the BHPS. The very high item non-response for self-employed earnings may be noted. This is found in many other surveys. In UKHLS, item non-response to the investment income question is followed-up with unfolding brackets (see section 3.2).

Figure 2: Item non-response rates for key income variables

Notes: General Population Sample
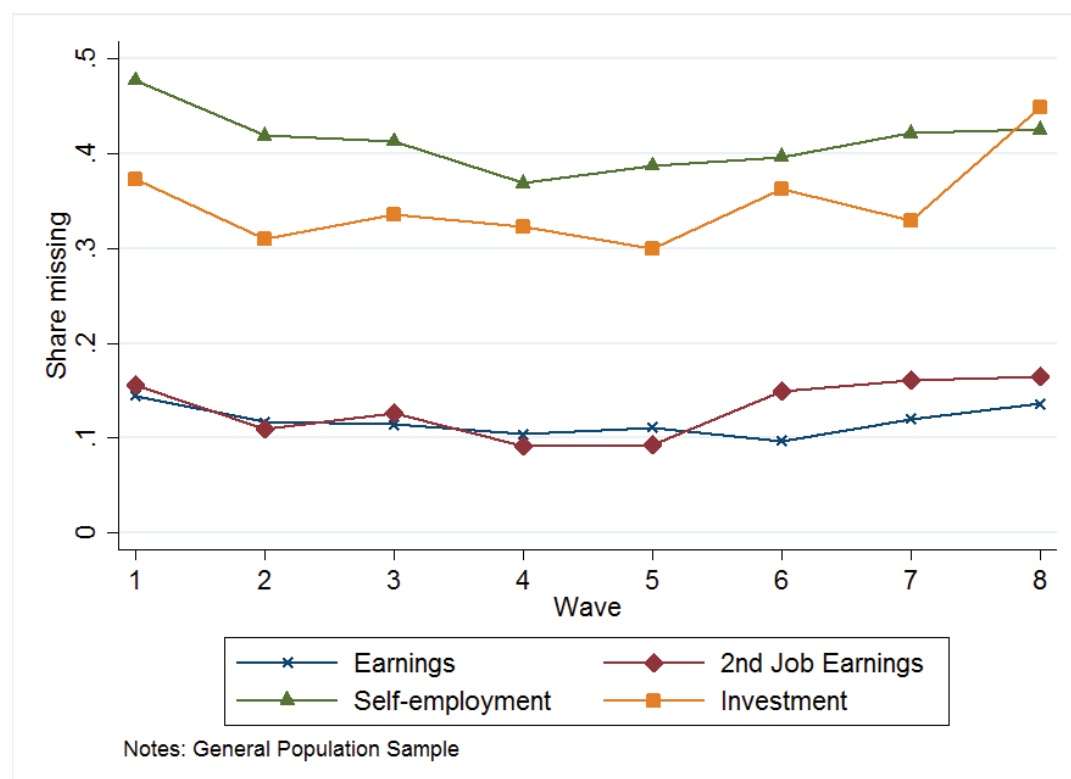
2.7.   WEIGHTING AND USE OF THE DIFFERENT SAMPLES

UKHLS has a complex sample design and is used in various ways by data analysts. Consequently, the weighting strategy is also complex. UKHLS provides weights for the household and individual levels, including all individuals enumerated in respondent households, and eligible individuals that respond or do not respond to different instruments, e.g. the adult questionnaire, the self-completion instrument, for responding to different combinations of study waves, and for the diverse sample components.

In general, weights are the product of a design weight to convey the probability of selection, adjustment for non-response, and sometimes post-stratification, to make the distribution a closer match to the population distribution.

Weights account for the fact that selection probabilities differ within sample component. For example, the members of different ethnic minority groups in the EMB and IEMB have different probabilities of selection. In addition, the countries in the former BHPS sample have different sampling fractions, including boost samples for Scotland, Wales and Northern Ireland.

Weights also account for the fact that selection probabilities differ across sample components, and thus can be used combine these components. For example, when combining the general population component with the former BHPS, the weights adjust for the fact that the BHPS sample does not contain immigrants for its period of fieldwork.

Finally, weights take the time pattern of response into consideration. For example, weights for complete longitudinal responses take into account differential probabilities of attrition after wave 1. They would include those for Waves 1 and 2 or Waves 1, 2, and 3. Cross-sectional weights are also produced. More information on the available weights can be found in Lynn and Kaminska (2010) and Knies (2017).

Longitudinal income analysis typically uses one of the adult individual longitudinal weights. However, if the concern is with the whole population including children, then the sample of enumerated individuals should be used with the enumerated individual weights. This is the case where a population level measure is of interest, for example a relative poverty line based on 60% of median equivalised income, then the enumerated individual weights should be used to compute median income. Jenkins (2011) provides a further discussion of the use of weights in income analysis.

3. INCOME DATA IN THE UKHLS

3.1. OVERVIEW

In this section we present an overview of the income data collected in UKHLS. It begins by describing the questionnaire content and emphasising key features of the income data collection. Next, it records questionnaire improvements that have been implemented over time. It also documents other issues relevant for income analysis including minor mistakes that slipped into the questionnaire script, that have since been corrected. It then briefly discusses new and exciting possibilities for administrative data linkage before describing how the UKHLS derived income variables are constructed.

3.2. QUESTIONNAIRE CONTENT

A significant proportion of questionnaire time is devoted to collection of income measures. The measures aim to cover all sources of individual incomes. Sources covered by the survey are: earnings from main and second jobs, social security benefits, state and private pensions, private transfers and investment income. Still, in the interests of limited questionnaire space, compared to surveys which specialise more on income, such as household budget surveys, compromises have been made.

UKHLS asks all adult respondents about each of the separate sources of income received, rather than asking for a global figure, which respondents will often not know and will report with significant error (Micklewright and Schnepf 2010). This approach also supports the analysis of separate income sources. For example, there are significant research agendas around using data on earnings, state benefit receipts, pension and other sources of income.

In UK income research and analysis, it is more common to use a measure of current income rather than annual income, which is more often used in other countries. Current income measurement is based on taking the usual amount received of an income source and converting it to a standard time metric. In the UKHLS case, all income receipts are converted to a monthly equivalent. Therefore, for employee earnings, for example, a usual amount from the employer is collected and the amount converted to a monthly equivalent. Where income is received more normally on an annual basis (e.g. income from savings and investments) it is asked on this basis and converted back to monthly.

The diverse topic coverage of Understanding Society leads to a questionnaire design with multiple modules defined by topic. The income questions are in different modules of the individual questionnaire. Earnings questions are asked as part of employee and self-employee modules; a

second jobs module collects data on gross earnings from second jobs; social security benefits, state and private pensions, and private transfers are collected in the "unearned income and state benefits module" and investment income as part of a "household finances" module. In contrast, housing costs and local tax measures are collected in the household questionnaire.

EMPLOYEE AND SELF-EMPLOYEE PAY

The employee's module asks employees to report their gross and net pay at last payment, the time period it covered, and their usual pay if "last" and "usual pay" differed. During the interview, respondents are encouraged to check a recent payslip to help them report more accurately and interviewers are instructed to probe for an approximate amount where a respondent cannot remember with the aim of reducing item non-response. Further checks are built into the questionnaire script to reduce mistakes that may occur during the interview. These checks occur where a respondent reports: a net income greater than gross, a net income equal to gross, or where net income differs substantially from gross income.

The self-employment module asks self-employees for their share of the profit or loss figure on their most recent accounts or, where this is not available, an estimate of their usual monthly or weekly income and whether this is before or after tax and national insurance contributions.

SECOND JOBS

The second job module asks all respondents if they have a second job and, if so, for the gross income in the last month from that job.

INVESTMENT INCOME

All respondents receive the household finances module which asks for income from interest and dividends in the last 12 months (to the nearest pound). To reduce missing data, where respondents cannot give an exact amount they are presented with a series of unfolding brackets where they can bound their annual investment income. Those reporting on unfolding brackets make up 7.3 percent of respondents reporting a valid investment income at wave 1. For individuals reporting on bounds, the Understanding Society survey team impute an exact amount of investment income (see section 4).

There are two stages to the collection of the 'unearned income and state benefits module'. The first stage collects information on whether a respondent is currently receiving each of 41 (originally 39) income sources; the second stage loops over each reported source to ask for the amount received. The 41 income sources covered by the module include the main UK state benefits; pensions (various including state, private and employer); family transfers and other income streams. An exact list of the 41 sources is included at the end of this document in table 15.

A novel feature of stage one of the module is that it finishes with a dependent interviewing loop. Dependent interviewing is a tool used in panel surveys that has been shown to reduce spurious change in reports between waves and thus improve data quality (Lynn, Jackle, et al. 2006). The dependent interviewing loop in Understanding Society asks respondents if they still receive sources they reported at wave t-1 but didn't report during the present interview.

During stage two, respondents are asked for each source for the last amount received or the approximate amount if they don't know, the period it covered, whether it was received jointly and the person number of the joint recipient and whether reported before or after tax (the latter only for pensions and since wave 5). The joint receipt questions operate with the aim of being able to avoid double counting of income streams in the same household.

Where within household non-response occurs (a panel member may refuse to take part, be away from home or be too ill, for example) then, with a respondent's permission, a proxy interview is conducted that is based on a much shorter and less detailed version of the main questionnaire. The proxy questionnaire asks for gross total income and earnings each in 14 bands (weekly or annual) and whether currently receiving each of 12 of the most widely received state benefits.

Individual and household totals of monthly income can be constructed from summing the individual income components across all sources and across all the members of a household (with the use of imputation and simulation). Such variables are in the public release files and their construction is discussed more fully below.

LOCAL TAXES

The UK local tax is council tax. Council tax is derived by assigning each residential property to one of eight bands based on the property value. The tax is set as a fixed amount for each band. In the UKHLS context, the household questionnaire collects data on the council tax band of households.

The tax liability corresponding to each band is set at the local authority level and there is substantial variation across authorities. Using information on the local authority in which the household is situated, a household can be linked to external data on council tax rates to give exact estimates of (gross) council tax. Derived council tax variables are available in the public release files.

Also available, to some low-income groups, is council tax reduction (formerly council tax benefit) - a means tested scheme that reduces council tax liability. This is collected in the unearned income section of the individual questionnaire as above and can be deducted from gross council tax to give net council tax liability. Although it is primarily thought of as a property tax, there are individual level discounts where only one person resides in a property (single person discount). This can be simulated based on demographic information also collected in the survey.

HOUSING COSTS

A group of housing wealth questions are reported by the household respondent annually. These are: tenure type (owed, mortgage, rented, rent –free); the identity of the owner if in the household; the current value of the home; last monthly mortgage payment; years left on mortgage; year first took ownership of the property; price the property was bought for; year that the mortgage began and type of mortgage. Additionally, at waves 1 (2009/10) and 5 (2013-14) respondents were asked for "the total amount secured against the property, including your mortgage and any other loans secured on the property". A spending section then asks for the last rent payment if the accommodation is rented and the net rent payment if receiving a rent subsidy (housing benefit). A point we return to is that the reported gross and net rent values can imply an amount of housing benefit that is inconsistent with the sum of the reported amounts in the individual questionnaires (in particular the unearned income module). Derived housing cost variables are available in the public release file and discussed in section 5.

3.3.   QUESTIONNAIRE CHANGES

From time to time it has been necessary to refine the questionnaire content to reflect developments in the tax and benefit system or, in rare cases, to correct mistakes that have crept into the questionnaire routing. Questionnaire changes have only been made where there was a strong case to improve data quality.

The introduction of two new UK states benefits (Universal Credit and Personal independence payments) meant that the original list of 39 "unearned income and state benefits" was extended to cover 41 sources from wave 4 (2012-2013).

A new question was asked from wave 5 to establish whether pensions were reported before or after tax. If applicable, respondents were asked if the "last payment" of pensions was before or after tax.

For wave 6 (2014-2015), stage one of the "unearned income and state benefits" section was redesigned to remove filtering questions. Filtering questions are used to reduce questionnaire time and respondent burden where it can be established that a group of questions do not apply to a particular respondent. In the case of "unearned income and state benefits", prior to wave 6 respondents were first asked about the receipt of broad types of payments (e.g., disability benefits, pensions, unemployment benefits) and then received follow-up questions asking about the receipt of specific payments only if they were relevant. The downside of this approach is that if the first filtering question does not work well, then income sources go unreported when they are in fact received.

There was some evidence of this occurring in the original questionnaire design and so it was decided that the filtering stage would be removed. This was implemented in the wave 6 questionnaire and so, from wave 6 onwards, all respondents receive all the specific benefit questions.

3.4.   KNOWN INCOME ISSUES

Section 6 will show that UKHLS estimates of the income distribution perform reasonably well, relative to a specialist income survey. However, as with all surveys, errors can occur during data collection. This subsection lists known issues in the data.

UKHLS uses dependent interviewing to reduce under-reporting of unearned income (see above), which is known to occur in household surveys. Overall, dependent interviewing improves data quality, however, one consequence is that it leads to differences in data quality between the baseline (wave one) interview and subsequent interviews where it takes effect (Fisher 2019). Mean income (in particular benefits and unearned income) has been shown to be lower and of lower quality at wave one, compared to later waves where it is functional. The causal effect of DI is large enough that it leads to statistically significant differences in population mean income between wave one and later waves.

A distinct issue that leads to data quality differences between wave one and later waves is panel conditioning. Panel conditioning occurs where respondents change their behaviour (either reporting or economic) as a result of being part of the panel. (Fisher 2019) shows that panel conditioning leads to higher reports of income, that correspond to higher quality data, from wave two onwards. The effect seems to be concentrated in 'benefits and unearned income' and is fairly large in magnitude. Related effects have been documented in other household surveys. Experienced panel members are known to be more likely to answer questions about their income in the US General Social Survey (Halpern-Manners, Warren and Torche 2014) and report higher incomes in the German Socio-Economic Panel (Frick and Grabka 2004).

A separate minor scripting error in the wave one interview occurred in the "unearned income and state benefits module" where respondents reporting receipt of either "other disability related benefits", "other regular payment" or "income from any other state benefit" were not asked for the amounts received on those benefits. These sources tend to be relatively minor and so the few missing amounts are filled by imputation by the data providers. The problem was fixed in the wave two questionnaire script.

A minor error in the questionnaire script meant that the hard checks on employee pay, described above, were not implemented before wave 6. The result is that a handful of respondents were allowed to report a net pay greater than the gross pay before wave 6. The data providers correct such cases by simulation where possible (see section 3.6 below).

The fieldwork agency responsible for data collection changed at Wave 6, while overall leadership of the study remained at the Institute for Social and Economic Research. The change is associated with a small reduction in overall response rates (section 2.6) and also a small increase in item non-response on the income variables. In practical terms, the effect on the UKHLS income series, if discernible, is small and by wave 7 response patterns had returned to their pre-change levels.

3.5. ADMINISTRATIVE DATA LINKAGE

UKHLS respondents were asked to consent to link their survey responses to administrative records on the receipt of state benefits held by the Department for Work and Pensions and tax records held by HM Revenue and Customs (HMRC). Income is known to be under-reported in surveys (eg. Brewer, Etheridge, O'Dea (2017); Lynn, Jackle et. al. (2012)) and so administrative data linkage offers an

exciting way forward to potentially reduce measurement error that may occur in survey reports of income.

64 percent of UKHLS respondents consented to link to their DWP records and 61 percent consented to link to their HMRC records. While the research potential is exciting, the processes and procedures for making linkage possible are detailed and complex, requiring carefully sought permissions from participants and the co-operation of external organisations who own the administrative datasets. Linkage has not yet been achieved but the data providers are currently implementing the linkage with the organisations in question.

Two notes of caution. First, some sources of income are simply not recorded in administrative records and so administrative data can supplement, but will never be able to completely replace the survey content eg. transfers between friends/family members are not recorded in administrative records. Second, a growing literature has documented errors in administrative data also (eg. Bollinger et al. (2018)) and the types and extent of such errors is only beginning to be understood.

3.6. CONSTRUCTION OF DERIVED INCOME VARIABLES

A range of income derived variables are available in the public release data. They are produced for each of the individual source variables in the questionnaire i.e., employee earnings, second job earnings, self-employee earnings, investment income and the 41 unearned income sources. Additionally, a set of measures of personal income and household income, both before and after taxes are available. The exact set of UKHLS income derived variables is set out in section 5 and summarised in table 13.

The construction of derived income variables in UKHLS broadly follows the approach used by Jenkins and his colleagues in constructing BHPS net incomes (with some differences discussed below). The approach has 8 steps:

1. Derive individual monthly amounts associated with each source. Figures 33-41 shows this graphically.
2. For labour income, where a net (gross) amount is provided, but the net (gross) amount is missing, then simulate the net (gross). See below.
3. For unearned income, identify payments that have been reported twice to avoid double counting of income e.g., both partners in a couple reported the same "child benefit". Details are in section 7.2.

4. Identify any obvious reporting errors in the data and set to missing. See below.

5. Impute missing income values (for respondents, proxies and non-respondents) and for gross and not net earnings. See section 4 for details.

6. Simulate remaining missing net labour income.

7. Sum complete data across income sources to get personal income totals.

8. Sum complete data across household members to get household income totals.

9. Calculate and implement other household deductions (council tax, housing costs).

The UKHLS simulation of missing net labour income is similar to that developed by Jenkins for the former BHPS and described in Jenkins (2011). Based on the reported or imputed gross income, income tax, national insurance and pension contributions are simulated, which are then deducted from gross labour income. In some cases, only respondents net amount is reported, and here gross can be similarly arrived at by using an iterative process. This UKHLS approach was developed in conjunction with DWP and the Child Poverty Unit and sought to replicate where possible the main income components used in the Household Below Average Income (HBAI).

The simulation proceeds as follows. First, the individual taxable income (labour income only) and the corresponding income tax are calculated. The income tax is calculated in line with the parameters of the tax system. These parameters are updated at each annual data release to reflect changes to the tax system. Then, using a similar approach, calculate national insurance contributions (NI) that are determined by gross earnings and NI rule and parameters, where differing rules are applied for employees and the self-employed. Third, employee pension contributions can be calculated from the reported rate of contributions from gross earnings, or, where this is missing, it is imputed at a fixed rate of gross earnings. Individual net labour income is then calculated as gross labour income minus the simulated tax, NI and pension contributions.

Reporting errors occasionally occur in household surveys which could lead to wrongly estimated amounts. For example, errors could arise from respondents confusing the period code, slipping decimal places, or reporting the wrong amount by mistake. In UKHLS, the hard checks, described above, aim to reduce such errors at the data collection stage. Remaining reporting errors should be rare. It is not possible to identify all reporting errors post data collection (at least in the absence of linked administrative records. See section 3.5). However, it is possible to identify obvious errors. For example, where a state benefit is known to be paid at fixed rate the reported data is edited to match the fixed rate. Cases reporting earnings that would place them in the top 0.1% of the distribution are subject to checks on occupation and implausible amounts are set to missing and imputed; similarly,

based on external sources, for each state benefit a maximum value is set above which the reported amounts are set to missing and imputed. To give an indication of the scale of editing, at wave seven, 13/20,821 observations on gross pay and 205/34,957 state benefits amounts were set to missing and imputed (see section 4).

3.7. EQUIVALISATION AND PRICE INDICES

Household needs grow with household size, but in a way that is not proportional. This means that a per person measure of income (ie. dividing household income by household size) gives a misleading picture of the living standards of each individual. In order to compare incomes for households of different size and composition, each net household income value should be adjusted by an equivalence scale. This is a standard adjustment that should be made when analysing survey data on household incomes. The equivalence scale included in the UKHLS public release files is the OECD-modified (Organisation for Economic Co-operation and Development) equivalence scale. It is a widely used equivalence scale that, for example, is adopted in official UK income distribution statistics (DWP 2018).

The OECD-modified scale gets a weight of one for the first adult in a household, 0.5 for each additional adult and 0.3 for each child. Total household income is then divided by the sum of the weightings to yield equivalised income. Note, differently from other countries, the official UK income distribution statistics are based on a normalisation of the scale so that a childless couple gets a weight of one (and not 1.5). In our validations of section 6, we do not perform this additional normalisation.

Various other equivalence scales have been proposed and it is beyond the scope of this paper to review them. We note only that they can lead to a different view of the relative living standards of different groups. For a review of different equivalence scales see Canberra Group (2011).

In order to compare incomes collected in different months and years, each net household income value should also be adjusted by a price index to a common time period. The price index should match the income concept understudy, that is, it should include items that can be purchased with that income. For example, the official UK income distribution statistics use different price indices for studying before and after housing costs incomes. Once a price index is chosen, it can be matched to a household's month of interview in each UKHLS survey year, so that the appropriate adjustment to incomes can be performed.

As with equivalence scales, many price indices are available including the Consumer Price Index (CPI) and Retail Price Index (RPI). For a detailed discussion of the use of price indices in household income analysis see Canberra Group (2011).

The price index we use in our validations of section 6 is a bespoke monthly CPI price index that is used in the official UK income distribution statistics and produced by the Office for National Statistics (see DWP (2018)). The index - Consumer Price Indices series excluding rents, maintenance repairs and water charges - is freely available for [download](#).

3.8.   DIFFERENCES WITH THE FORMER BHPS

While the former BHPS is not the focus of the present paper, for the interested reader, we describe how its income variables differ from the UKHLS ones. While the income concepts in each survey are broadly comparable, there are differences in points of detail. Where possible, harmonisation of income variables across the surveys has taken place and details can be found in Fumagalli, Knies and Buck (2017). Detailed information on the derivation and quality of the BHPS income variables has been carefully documented elsewhere including Levy and Jenkins (2012) and Jenkins (2011).

The differences between the surveys arise from differences in the questionnaire content and the variable construction. In most cases, the changes reflect an improvement in the UKHLS methods relative to the BHPS ones. In what follows in this subsection, we refer to the BHPS prior to its entry in the UKHLS as "BHPS" (not to be confused with the "BHPS sample component" of UKHLS whose content and income variables are identical to those in the other UKHLS subsamples).

Both UKHLS and BHPS derive measures of household and personal income, net and gross, and by income subcomponent. Both surveys aim to provide the same derived variables with the exception of annual income. In the BHPS, both annual and current measures were produced. Greater pressure on questionnaire space in the UKHLS meant that it was not possible retain the questions required for both measures and the choice was made to prioritise current income, so that there are no annual measures of household income.

The two biggest sources of difference relate to imputation. In the UKHLS, derived household income variables are produced for households which contain adult non-respondents, whereas in the BHPS they were not. This means that analysis based on the UKHLS household income variables essentially adjusts for within household non-response. This issue has become more important since the early days of the BHPS (1990's) when household response rates were higher and analysis of complete

households only was a lesser threat to representativeness (see subsection 2.6). Second, the BHPS used a longitudinal imputation based on a moving three wave window whereas UKHLS moved to a method based on the Little and Su method. This method is used in other international panel surveys (eg. HILDA, SOEP, SHP) and evidence suggests that it outperforms alternatives (Starick and Watson 2009). Both these changes should therefore represent an improvement over BHPS imputation methods. Full details on the UKHLS imputation methods are provided in section 4.

Two final (minor) differences in definitions and construction are not expected to lead to large differences between the series. First, the questionnaires differ slightly. The content on earnings from main and second jobs is identical in both surveys. Similarly, both surveys collect the same range of unearned sources, but the structure differs slightly where BHPS had an initial question and show cards that differ from the UKHLS structure described above. The annual investment income question is identical in both surveys, from wave 9 of BHPS: the exact amount of investment income is asked, and then unfolding brackets are used to collect information for item non-respondents. Before BHPS wave 9, the question wording was identical, but the answers were directly collected in three reporting bands. Second, while the approaches to simulation of net incomes are largely similar they are not identical. In both surveys, the calculations focus on labour income (first and second jobs). But the UKHLS estimates have sought to include tax deductions from non-pay income (rental income) which are assumed to be reported gross, and we believe this was not done in BHPS.

Two final points relate to differences around the UKHLS-BHPS seam. First, as described above, while interviewing occurs at an annual frequency in both surveys, the gap between interviews at the seam is longer. The gap between the first UKHLS interview and last BHPS ranges from 16-27 months depending on the interview allocation date. This larger gap is expected to lead to more income changes across the seam relative to what is seen between other waves. Second, there was a drop-off in response rates between wave 18 of the BHPS and wave 1 of UKHLS. Of all persons who completed an individual interview at wave 18 of BHPS, 79.4% also completed the Understanding Society wave 2 interview (Lynn, Burton, et al. 2012). The comparable figure for other BHPS wave pairs (Brice, Buck and Prentice-Lane 1993) ranges from 86% (between wave 1 and 2) and 95% (between wave 17 and 18). In principle, weighting analysis should account for this response change. There is evidence that, at least for average earnings, the UKHLS-BHPS series performs well across the seam. Postel-Vinayy and Sepahsalari (2019) have shown that, across the seam, a UKHLS-BHPS average earnings series closely matches an equivalent series derived from ONS sources (the authors reach the same conclusion for UKHLS-BHPS series of aggregate worker stocks and worker flows across the seam).

4. IMPUTATION OF MISSING DATA

4.1. OVERVIEW OF THE IMPUTATION PROCESS

Figure 3 shows how missing income information arises, as well as the type of income information we impute and release. No income data are imputed nor released for non-responding households, i.e. for cases where there is no household questionnaire nor there are individual questionnaire. If we have some information about the household, for example the household questionnaire, the household is considered a responding household. All incomes of all individuals in responding households are imputed and used in the creation of personal and income aggregated. However, only some of these sources are released, as shown in Figure 3.

Figure 3. What is imputed and released
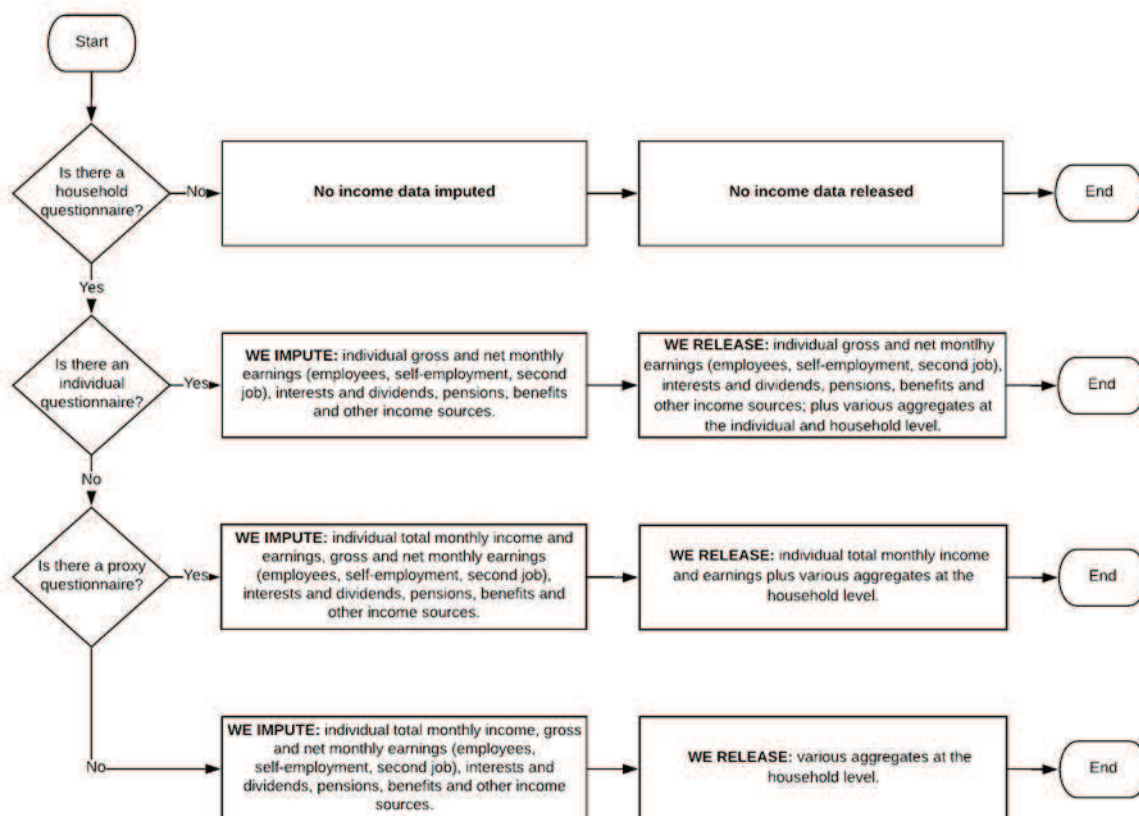
The imputation of missing income data is done in two steps. The first step is the "initialisation": missing values are imputed using cross-sectional imputation (deterministic imputation, stochastic imputation, stochastic imputation via chained equations, and hot deck) and longitudinal imputation

(carryover methods). This permits to obtain a rectangular dataset where all individuals in responding households have a reported or imputed value for each income source. In the second step, the values imputed in the first step are replaced using the longitudinal imputation method by Little and Su (1989). The Little and Su method is widely used in other major longitudinal surveys, such as the Australian household panel- HILDA-, the Longitudinal Study of Australian Children –LSAC-, the German Socio-Economic Panel –SOEP- and the Swiss Household Panel –SHP (Frick and Grabka, Missing Income Data in the German SOEP: Incidence, imputation and its Impact on the Income distribution 2004) (Frick and Grabka 2007), (Hayes and Watson 2009), (Starick and Watson 2009), (Lipps 2010), (Mullan, Daraganova and Baker 2015). The imputation process is summarised in Figure 4.

For individual respondents in responding households, the initialisation is carried out by year-wave using a range of parametric and semi-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 method used is predictive mean matching imputation (PMM). The income variables and their predictors are generally imputed jointly, using chained equations (ICE). A summary of the characteristics of the initialisation process for responding individuals is reported in table 5.

For proxy respondents and non-respondents in responding households, the initialisation is carried out using longitudinal carryover methods and hot-deck. Carryover methods are used to impute employment status. Income sources are imputed by hot-deck. Missing values in the variables defining the categories for the hot deck –other than employment status- are set equal to their median. Note that the proxy questionnaire only collects information on total personal income and earnings reported in bands. There is no income information for individual non-respondents. Therefore, neither proxy respondents nor individual non-respondents have any reported income sources. This means that all income sources need to be imputed. The hot-deck method permits to impute all income sources by taking them from the same donor. For proxy respondents, this makes sure imputed income sources are coherent with the reported bounds. A summary of the characteristics of the initialisation process for proxy respondents and non-respondents in responding households is reported in table 6.

Figure 4. Imputation summary
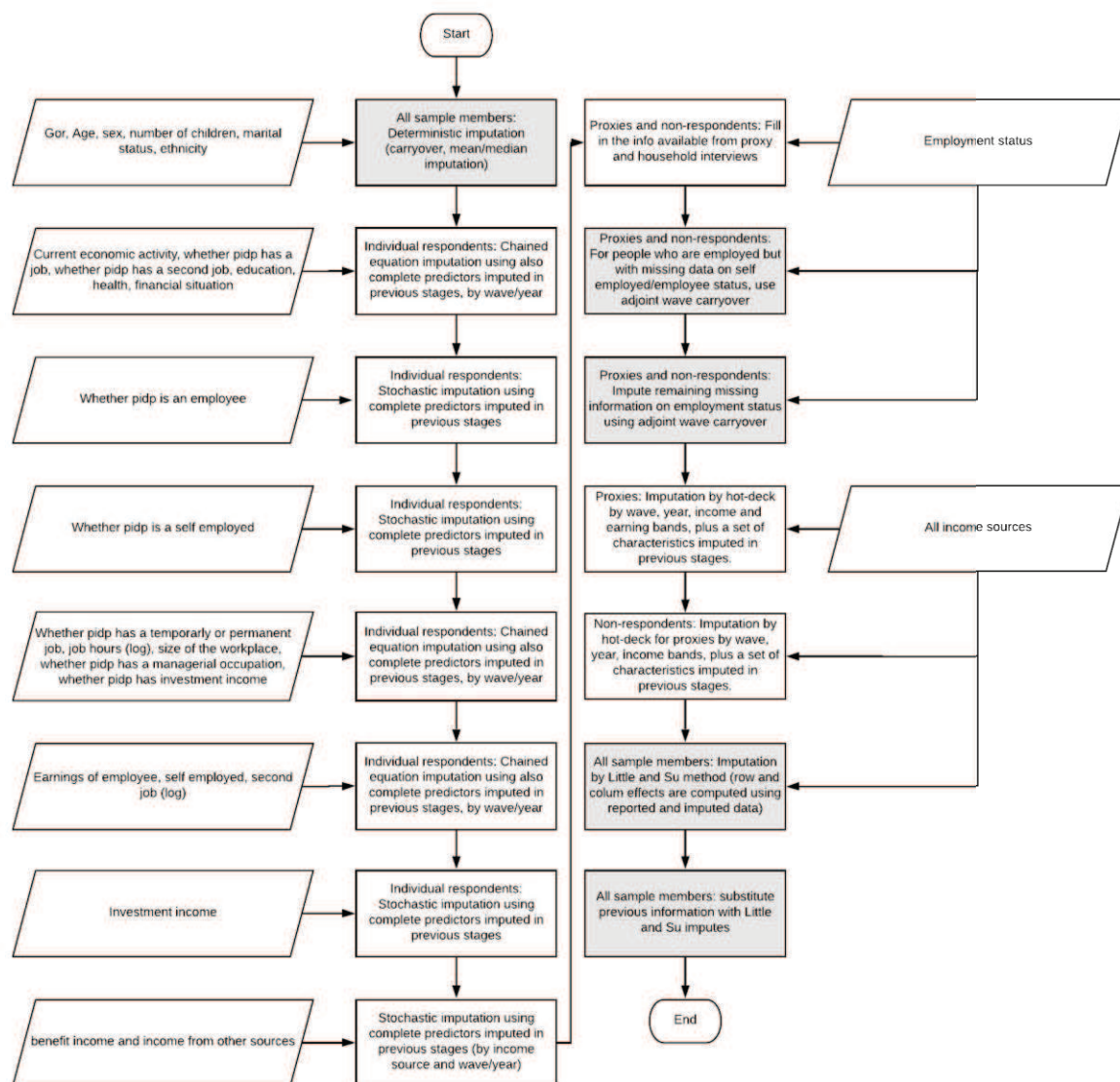
Note: grey boxes indicate the presence of some form of longitudinal imputation

Table 5. Initialisation for responding individuals

Variable name	Label	Imputation method	Predictors (regressors)
paygu_dv	usual gross pay per month: current job	linear regression, through chained equation	Education, region, sex, age (+square), number of children (+square), marital status, ethnicity, health, whether the job is permanent or temporary, job hours (log), size of company, soc classification, managerial duties, earnings from second job
seearngrs_dv	self employment earnings - gross	linear regression, through chained equation	Education, region, sex, age (+square), number of children (+square), marital status, ethnicity, health, whether the job is permanent or temporary, job hours (log), size of company, SOC classification, managerial duties, earnings from second job
j2pay_dv	pay in second job	linear regression, through chained equation	Education, region, sex, age (+square), number of children (+square), marital status, ethnicity, health, whether the job is permanent or temporary, job hours (log), size of company, SOC classification, managerial duties, earnings from first job (either from self employment of as an employee)
fiyrinvinc_dv	income from savings and investments, annual	Whether pidp has investment income: predictive mean matching	Education, region, sex, age (+square), number of children (+square), marital status, ethnicity, whether pidp is an employee, whether the individual is a self employed, financial situation
		Amount of investment income when positive (in logs): interval regression	Education, region, sex, age (+square), number of children (+square), marital status, ethnicity, whether pidp has a job, whether the first job is temporary or permanent, job hours, company size, SOC classification, earnings from first and second job, whether the individual has a second job, financial situation
frmnthimp_dv	monthly income received from benefit/other miscellaneous income sources	Predictive mean matching or interval regression when missing is only partial	Education, region, sex, age (+square), number of children (+square), marital status, ethnicity, health, financial situation, sum of earnings and investment income

Table 6. Initialisation for proxy respondents and individual non-respondents in responding households

Variable name	Label	Proxy respondents		Non respondents	
		Imputation method	Predictors (categories)	Imputation method	Predictors (categories)
paygu_dv	usual gross pay per month: current job	Hot-deck	Total income bands, total earnings bands, year, wave, sex, age, education employment status, sample origin, marital status, whether pidp is a parent, housing tenure, health, number of durables	Hot-deck	Year, wave, sex, age, sample origin, marital status, whether pidp is a parent, housing tenure, number of durables.
seearngrs_dv	self employment earnings - gross				
j2pay_dv	pay in second job				
fiyrinvinc_dv	income from savings and investments, annual				
frmnthimp_dv	Total income from benefit/other miscellaneous income				

Note: When a donor matching the full set of categories is not found, categories are made coarser, or, ultimately, removed from the list of predictors.

4.2. IMPUTATION PROCEDURES

IMPUTATION BY CHAINED EQUATIONS (ICE)

Imputation by chained equations is a multivariate stochastic imputation method used to impute a set of variables jointly.[7] ICE allows for interdependence between the imputed variables by estimating each variable sequentially (Van Buuren, Boshuizen and Knook 1999), (Ragunathan, et al. 2001). The ICE method has been used in major household panel surveys such as the ECHP, as well as in combination with the Little and Su method (Westermeier and Grabka 2016).

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

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

---

[7] For more details about stochastic imputation, see Rubin (1987) , Schafer (1997), and Kenward and Carpenter (2007)

$Y_1, Y_2, ..., Y_k$ are the income and auxiliary variables to be imputed, ordered from the one with the smallest percentage of missing values, Y_1, to the one with the largest percentage of missing values Y_k. X is a set of auxiliary variables observed for all individuals, $\alpha_{10}, ..., \alpha_{ko}, ..., \alpha_{kk-1}$ and $\beta_{10}, ..., \beta_{ko}, ..., \beta_{kk-1}$ are parameters, and $u_1, u_2, ..., u_k$ are random errors. This recursive system allows us to impute each variable separately and sequentially in the following steps:

i.   The first equation is estimated, and the missing values for Y_1 are imputed.

ii.   The second equation is estimated by replacing the missing values of Y_1 with those imputed in the previous step. The missing values of Y_2 are imputed.

iii. The above steps are repeated sequentially for each of the remaining equations, until all missing values in $Y_1, Y_2, ..., Y_k$ have been imputed.[8]

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, using as explanatory variables both X and the imputed variables $\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_k$, except the one used as dependent variable. At the end of this new cycle, a new set of imputed variables is produced and used to begin a further cycle of imputations. These cycles of imputations are repeated until convergence.

PREDICTIVE MEAN MATCHING (PMM)

Predictive mean matching is a semi-parametric imputation method. 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 (R. Little 1988). This is done in four steps:

i.   Regression models for the variable to be imputed are estimated

ii.   Fitted values are produced

iii. Records with missing information (recipients) are matched to donors based on the fitted values computed in ii)

iv.   Missing values are replaced with observed values from donors.

---

[8] Imputed values are drawn from the posterior predictive distribution of the variable to be imputed, conditional on the observed data. Note that not all the X and the Y predictors need to be included in every equation (see table 6 for the list of the predictors used in the imputation of our income data).

HOTDECK (HD)

Hot-deck is a non-parametric imputation method. 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 (see table 3). Once a suitable donor is identified, information on all income sources are carried over from the donor.

POPULATION CARRYOVER (PC)

PC is a longitudinal imputation method. It 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

LITTLE AND SU (LS)

The Little and Su method is a longitudinal non-parametric imputation method. The version of the LS method used in Understanding Society imputes missing values using a multiplicative model (Little and Su 1989). For any given income component, the final imputation is the product of 3 terms: a row (unit) effect, proportional to the person-specific mean of that income component, a column (time) effect proportional to its time-specific mean, and a residual effect donated from another respondent with complete information for that income component.

In Understanding Society, we initialise the Little and Su by using a combination of cross sectional and longitudinal methods to impute all income sources for all individuals in a responding household (see previous sections). Then, the LS method is applied as follows.

First, the data are organised in wide form. Table 7 shows a simple hypothetical case of earnings data after the initialisation but before going through the LS imputation.

Table 7. Example of earnings data after the initialisation phase

pidp	Earnings w1	w2	w3
1	135	130	200
2	200	200 *	200 *
3	200	480	210
4	380 *	400	420
5	350	370	300 *
6	235	243	342
7	400 *	450	470
8	0	790	790
9	0 *	790	790
10	360	450	600
11	675	235	700
12	345	690	800
13	.	230 *	0
14	0	230 *	0

Note: *indicates imputed data. The zeros indicate cases where the individual was not applicable for earnings data, for example due to unemployment.

Note cases of interest 9, 13, and 14 which correspond to an imputed inapplicable, a temporary sample member (ineligible at wave 1), and a respondent applicable only in one wave.[9]

Second, the column effects are calculated. In our case, column effects are cross-sectional mean earnings at wave x as a proportion of mean earnings across all waves. For example, a column effect of 0.7 at wave 1 means that wave 1 mean income is 70% of overall mean income.

In Understanding Society, column effects are calculated based on cases that are applicable in all waves (either reported or in the initial imputes). In our case, pidps 8, 9, 14 are excluded as inapplicable in some waves, while 13 is excluded as it is ineligible at wave 1.[10]

Third, row effects are calculated. Row effects are mean earnings for each individual where each reported value is scaled by the individual's column effect. When calculating the row effects, we make use of the imputes obtained in the initialisation phase. Waves in which a respondent is

[9] For any given income source, are considered "applicable" respondents who should have a value for that particular income source. For example unemployed are inapplicable for both employee and self-employment earnings. Self-employed are applicable for self-employment earnings, but not for employee earnings.
[10] For any given income source, are considered "eligible" respondents for whom we have an individual or proxy questionnaire, in other words, anyone who is not a non-respondent.

inapplicable are not included in the calculation of the row effects. For example, in the case of pidps 8 and 9, only the –rescaled – non-zero values at waves 2 and 3 are used. Table 8 shows row and column effects for the data in table 7.

Table 8. Example of earnings data: row and column effects.

		Earnings		
pidp	w1	w2	w3	Row effect
1	135	130	200	153.8
2	200	200 *	200 *	202.2
3	200	480	210	300.4
4	380 *	400	420	402.8
5	350	370	300 *	346.1
6	235	243	342	271.7
7	400 *	450	470	442
8	0	790	790	749.9
9	0 *	790	790	749.9
10	360	450	600	464.9
11	675	235	700	540.2
12	345	690	800	599.4
13	.	230 *	0	234.7
14	0	230 *	0	234.7
Sum	3280	3648	4242	
Column Effect	0.8809311	0.97977	1.1393	

Note: the average of the sum of earnings is 3723.3

Fourth, data are sorted by row effects.

Fifth, imputed data from the initialisation process are substituted with LS imputes. This is done as follows. For each pidp with imputed income data, we identify the closest donor, i.e. the individual with the closest row effect. Then, we replace all imputed data with the corresponding donor value multiplied by the ratio of the recipient-donor row effects. Note that a single donor is used for each pidp with earnings missing at least one wave. Donors should therefore be applicable and non-imputed in all waves for which they are donating (see cases 13, 14, 8, 9 who do not qualify as donors). Table 9 shows LS imputes for the data in table 7.

Some cases do not get a LS impute. They are:

i. The non-applicable cases

ii. Individuals applicable at only one wave (cases 13, 14 and 9).

iii.   Cases for which the row effect of the donor and the row effect of the recipient differ by more than 10%.

iv.   State benefit income because we want to impute amounts consistent with the rules of the benefit system (the cross-sectional PMM imputes aim to do that)

Table 9. Example of earnings data: row and column effects, plus LS imputes.

pidp	w1	w2	w3	Row effect
			Earnings	
pidp	w1	w2	w3	
1	135	130	200	153.8
2	200	170.912 *	262.94 *	202.2
13	.	230 *	0	234.7
14	0	230 *	0	234.7
6	235	243	342	271.7
3	200	480	210	300.4
5	350	370	241.92 *	346.1
4	407.31 *	400	420	402.8
7	342.27 *	450	470	442
10	360	450	600	464.9
11	675	235	700	540.2
12	345	690	800	599.4
8	0	790	790	749.9
9	0 *	790	790	749.9
Column Effect	0.8809311	0.97977	1.1393	

Note: data are sorted by row effect.

4.3.   EFFECT OF IMPUTATION

This section presents descriptive statistics on the effect of the imputation on the main income variables. Table 10 shows the share of income data by decile for the main income components, i.e., monthly gross earnings from the first job (for employees and self-employed, respectively), monthly gross earnings from second job, annual income from interests and dividends.[11] Table 11 reports the same information for the main income aggregates, i.e., total individual income from benefits and other sources, total individual gross labour income, total individual gross income, and total household gross income. Finally, for the main income components, table 12 compares descriptive statistics for the imputed and the non-imputed data.

---

[11] Deciles are computed on the full distribution, including both reported and imputed data.

Table 10: Share of missing data by wave and decile (main income components)

Decile	Wave 2			Wave 3			Wave 4			Wave 5			Wave 6			Wave 7			Wave 8		
	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N
Monthly gross earnings from first job (employees)																					
1	0.07	0.26	2555	0.07	0.25	2289	0.06	0.24	2107	0.06	0.25	2019	0.12	0.32	2015	0.08	0.35	1936	0.10	0.31	1829
2	0.08	0.27	2394	0.07	0.26	2183	0.07	0.25	2124	0.06	0.25	2017	0.11	0.31	2019	0.09	0.29	1939	0.10	0.30	1840
3	0.07	0.26	2471	0.07	0.26	2232	0.06	0.24	2095	0.08	0.26	2106	0.12	0.32	2094	0.08	0.27	1940	0.11	0.31	1820
4	0.06	0.25	2473	0.06	0.23	2428	0.05	0.23	2099	0.05	0.22	1962	0.12	0.32	1932	0.08	0.27	2052	0.09	0.29	1838
5	0.08	0.28	2474	0.08	0.28	2043	0.06	0.24	2289	0.06	0.24	2069	0.12	0.32	2036	0.08	0.27	1863	0.10	0.30	1894
6	0.08	0.27	2495	0.09	0.28	2233	0.07	0.25	1949	0.07	0.26	1934	0.11	0.31	1993	0.08	0.27	2105	0.10	0.31	1757
7	0.08	0.28	2451	0.07	0.26	2231	0.07	0.25	2089	0.07	0.26	2017	0.11	0.31	2015	0.08	0.27	1722	0.10	0.30	1824
8	0.07	0.26	2473	0.08	0.28	2234	0.08	0.26	2141	0.07	0.26	2031	0.11	0.31	2038	0.09	0.29	1908	0.09	0.29	1859
9	0.08	0.27	2679	0.07	0.26	2295	0.07	0.26	2082	0.08	0.27	2004	0.12	0.32	2035	0.08	0.28	1935	0.10	0.30	1799
10	0.08	0.28	2267	0.08	0.26	2173	0.07	0.25	2086	0.06	0.25	2016	0.11	0.32	1971	0.07	0.25	1931	0.10	0.30	1828
Monthly gross earnings from first job (self employed)																					
1	0.32	0.47	361	0.22	0.42	342	0.22	0.41	329	0.24	0.43	324	0.35	0.48	339	0.24	0.68	321	0.28	0.45	314
2	0.29	0.46	361	0.27	0.44	342	0.25	0.44	327	0.24	0.43	315	0.32	0.47	332	0.36	0.48	321	0.41	0.49	296
3	0.38	0.48	360	0.38	0.48	341	0.33	0.47	329	0.37	0.48	319	0.45	0.50	335	0.40	0.49	321	0.38	0.49	293
4	0.43	0.50	378	0.39	0.49	341	0.34	0.48	328	0.42	0.49	323	0.48	0.50	335	0.40	0.49	321	0.46	0.50	300
5	0.39	0.49	343	0.46	0.50	342	0.39	0.49	327	0.40	0.49	340	0.55	0.50	335	0.48	0.50	320	0.44	0.50	302
6	0.47	0.50	363	0.41	0.49	341	0.43	0.50	328	0.44	0.50	297	0.47	0.50	335	0.55	0.50	321	0.52	0.50	302
7	0.41	0.49	359	0.44	0.50	342	0.43	0.50	328	0.43	0.50	317	0.53	0.50	335	0.48	0.50	320	0.51	0.50	298
8	0.44	0.50	360	0.37	0.48	341	0.41	0.49	328	0.41	0.49	326	0.47	0.50	335	0.48	0.50	321	0.48	0.50	303
9	0.35	0.48	362	0.43	0.50	345	0.45	0.50	328	0.42	0.49	312	0.48	0.50	335	0.44	0.50	321	0.44	0.50	301
10	0.33	0.47	359	0.37	0.48	338	0.35	0.48	327	0.32	0.47	319	0.37	0.48	335	0.46	0.50	320	0.39	0.49	298
Monthly gross earnings from second job (self employed)																					
1	0.06	0.24	274	0.09	0.29	227	0.06	0.24	239	0.07	0.26	188	0.14	0.35	274	0.15	0.36	204	0.14	0.34	262
2	0.10	0.29	240	0.11	0.31	186	0.09	0.28	192	0.07	0.25	182	0.13	0.34	211	0.13	0.34	231	0.14	0.35	138
3	0.08	0.28	306	0.07	0.25	247	0.04	0.20	242	0.07	0.26	200	0.16	0.37	270	0.15	0.36	133	0.14	0.35	182
4	0.09	0.29	265	0.12	0.32	189	0.09	0.29	204	0.09	0.29	181	0.23	0.42	213	0.19	0.39	202	0.13	0.34	194
5	0.11	0.31	274	0.12	0.33	205	0.07	0.26	213	0.11	0.31	209	0.14	0.35	248	0.13	0.34	179	0.21	0.41	194
6	0.17	0.38	141	0.14	0.35	251	0.11	0.32	266	0.08	0.28	146	0.21	0.41	178	0.15	0.36	215	0.19	0.40	239
7	0.10	0.30	238	0.15	0.36	187	0.12	0.32	185	0.10	0.30	257	0.14	0.35	274	0.18	0.39	160	0.13	0.34	164
8	0.13	0.33	270	0.20	0.40	143	0.16	0.37	155	0.11	0.32	131	0.18	0.39	192	0.18	0.38	193	0.18	0.38	210
9	0.18	0.38	226	0.14	0.35	205	0.14	0.35	222	0.12	0.33	165	0.18	0.38	258	0.17	0.38	204	0.23	0.42	196
10	0.13	0.33	248	0.13	0.34	203	0.10	0.31	201	0.09	0.29	184	0.22	0.41	204	0.17	0.38	170	0.21	0.41	161
Annual income from interests and dividends																					
1	0.14	0.34	2162	0.17	0.37	1772	0.16	0.37	1907	0.14	0.35	1728	0.23	0.42	1531	0.20	0.40	1536	0.36	0.48	1446
2	0.21	0.41	999	0.25	0.43	946	0.26	0.44	948	0.21	0.41	808	0.31	0.46	1404	0.30	0.46	920	0.30	0.46	1492
3	0.22	0.42	2175	0.24	0.43	1793	0.22	0.41	1777	0.21	0.41	1528	0.32	0.47	1142	0.22	0.41	1236	0.37	0.48	1809
4	0.39	0.49	798	0.51	0.50	912	0.26	0.44	1895	0.24	0.43	1596	0.26	0.44	1610	0.26	0.44	1526	0.55	0.50	1032
5	0.25	0.43	1534	0.20	0.40	1358	0.40	0.49	713	0.38	0.49	662	0.33	0.47	1782	0.40	0.49	744	0.35	0.48	1444
6	0.25	0.43	1693	0.26	0.44	1350	0.30	0.46	1436	0.28	0.45	1264	0.37	0.48	1133	0.29	0.45	1193	0.40	0.49	1446
7	0.40	0.49	1374	0.36	0.48	1380	0.35	0.48	1317	0.37	0.48	1265	0.38	0.48	1291	0.27	0.45	1510	0.40	0.49	1444
8	0.29	0.46	1604	0.36	0.48	1328	0.35	0.48	1425	0.28	0.45	1264	0.40	0.49	1271	0.33	0.47	1139	0.47	0.50	1444
9	0.29	0.46	1546	0.30	0.46	1355	0.32	0.47	1428	0.26	0.44	1265	0.42	0.49	1066	0.36	0.48	950	0.40	0.49	1539
10	0.32	0.47	1449	0.34	0.47	1354	0.32	0.47	1426	0.30	0.46	1263	0.29	0.45	1358	0.29	0.46	1170	0.39	0.49	1350

Note: For earnings from first job, the share of missing cases is computed for applicable cases only. For earnings from second job and earnings from interests and dividends, the share of missing cases is computed for positive cases only. IEMB and wave 1 are excluded.

Table 11: Share of missing data by wave and decile (main income aggregates)

	Wave 2			Wave 3			Wave 4			Wave 5			Wave 6			Wave 7			Wave 8		
Decile	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N	Mean	sd	N
	Total individual income from benefits and other sources																				
1	0.07	0.26	2555	0.07	0.25	2289	0.06	0.24	2107	0.06	0.25	2019	0.12	0.32	2015	0.08	0.35	1936	0.10	0.31	1829
2	0.08	0.27	2394	0.07	0.26	2183	0.07	0.25	2124	0.06	0.25	2017	0.11	0.31	2019	0.09	0.29	1939	0.10	0.30	1840
3	0.07	0.26	2471	0.07	0.26	2232	0.06	0.24	2095	0.08	0.26	2106	0.12	0.32	2094	0.08	0.27	1940	0.11	0.31	1820
4	0.06	0.25	2473	0.06	0.23	2428	0.05	0.23	2099	0.05	0.22	1962	0.12	0.32	1932	0.08	0.27	2052	0.09	0.29	1838
5	0.08	0.28	2474	0.08	0.28	2043	0.06	0.24	2289	0.06	0.24	2069	0.12	0.32	2036	0.08	0.27	1863	0.10	0.30	1894
6	0.08	0.27	2495	0.09	0.28	2233	0.07	0.25	1949	0.07	0.26	1934	0.11	0.31	1993	0.08	0.27	2105	0.10	0.31	1757
7	0.08	0.28	2451	0.07	0.26	2231	0.07	0.25	2089	0.07	0.26	2017	0.11	0.31	2015	0.08	0.27	1722	0.10	0.30	1824
8	0.07	0.26	2473	0.08	0.28	2234	0.08	0.26	2141	0.07	0.26	2031	0.11	0.31	2038	0.09	0.29	1908	0.09	0.29	1859
9	0.08	0.27	2679	0.07	0.26	2295	0.07	0.26	2082	0.08	0.27	2004	0.12	0.32	2035	0.08	0.28	1935	0.10	0.30	1799
10	0.08	0.28	2267	0.08	0.26	2173	0.07	0.25	2086	0.06	0.25	2016	0.11	0.32	1971	0.07	0.25	1931	0.10	0.30	1828
	Total individual gross labour income																				
1	0.32	0.47	361	0.22	0.42	342	0.22	0.41	329	0.24	0.43	324	0.35	0.48	339	0.24	0.68	321	0.28	0.45	314
2	0.29	0.46	361	0.27	0.44	342	0.25	0.44	327	0.24	0.43	315	0.32	0.47	332	0.36	0.48	321	0.41	0.49	296
3	0.38	0.48	360	0.38	0.48	341	0.33	0.47	329	0.37	0.48	319	0.45	0.50	335	0.40	0.49	321	0.38	0.49	293
4	0.43	0.50	378	0.39	0.49	341	0.34	0.48	328	0.42	0.49	323	0.48	0.50	335	0.40	0.49	321	0.46	0.50	300
5	0.39	0.49	343	0.46	0.50	342	0.39	0.49	327	0.40	0.49	340	0.55	0.50	335	0.48	0.50	320	0.44	0.50	302
6	0.47	0.50	363	0.41	0.49	341	0.43	0.50	328	0.44	0.50	297	0.47	0.50	335	0.55	0.50	321	0.52	0.50	302
7	0.41	0.49	359	0.44	0.50	342	0.43	0.50	328	0.43	0.50	317	0.53	0.50	335	0.48	0.50	320	0.51	0.50	298
8	0.44	0.50	360	0.37	0.48	341	0.41	0.49	328	0.41	0.49	326	0.47	0.50	335	0.48	0.50	321	0.48	0.50	301
9	0.35	0.48	362	0.43	0.50	345	0.45	0.50	328	0.42	0.49	312	0.48	0.50	335	0.44	0.50	321	0.44	0.50	301
10	0.33	0.47	359	0.37	0.48	338	0.35	0.48	327	0.32	0.47	319	0.37	0.48	335	0.46	0.50	320	0.39	0.49	298
	Total individual gross income																				
1	0.06	0.24	274	0.09	0.29	227	0.06	0.24	239	0.07	0.26	188	0.14	0.35	274	0.15	0.36	204	0.14	0.34	262
2	0.10	0.29	240	0.11	0.31	186	0.09	0.28	192	0.07	0.25	182	0.13	0.34	211	0.13	0.34	231	0.14	0.35	138
3	0.08	0.28	306	0.07	0.25	247	0.04	0.20	242	0.07	0.26	200	0.16	0.37	270	0.15	0.36	133	0.14	0.35	182
4	0.09	0.29	265	0.12	0.32	189	0.09	0.29	204	0.09	0.29	181	0.23	0.42	213	0.19	0.39	202	0.13	0.34	194
5	0.11	0.31	274	0.12	0.33	205	0.07	0.26	213	0.11	0.31	209	0.14	0.35	248	0.13	0.34	179	0.21	0.41	194
6	0.17	0.38	141	0.14	0.35	251	0.11	0.32	266	0.08	0.28	146	0.21	0.41	178	0.15	0.36	215	0.19	0.40	239
7	0.10	0.30	238	0.15	0.36	187	0.12	0.32	185	0.10	0.30	257	0.14	0.35	274	0.18	0.39	160	0.13	0.34	164
8	0.13	0.33	270	0.20	0.40	143	0.16	0.37	155	0.11	0.32	131	0.18	0.39	192	0.18	0.38	193	0.18	0.38	210
9	0.18	0.38	226	0.14	0.35	205	0.14	0.35	222	0.12	0.33	165	0.18	0.38	258	0.17	0.38	204	0.23	0.42	196
10	0.13	0.33	248	0.13	0.34	203	0.10	0.31	201	0.09	0.29	184	0.22	0.41	204	0.17	0.38	170	0.21	0.41	161
	Total household gross income																				
1	0.14	0.34	2162	0.17	0.37	1772	0.16	0.37	1907	0.14	0.35	1728	0.23	0.42	1531	0.20	0.40	1536	0.36	0.48	1446
2	0.21	0.41	999	0.25	0.43	946	0.26	0.44	948	0.21	0.41	808	0.31	0.46	1404	0.30	0.46	920	0.30	0.46	1492
3	0.22	0.42	2175	0.24	0.43	1793	0.22	0.41	1777	0.21	0.41	1528	0.32	0.47	1142	0.22	0.41	1236	0.37	0.48	1809
4	0.39	0.49	798	0.51	0.50	912	0.26	0.44	1895	0.24	0.43	1596	0.26	0.44	1610	0.26	0.44	1526	0.55	0.50	1032
5	0.25	0.43	1534	0.20	0.40	1358	0.40	0.49	713	0.38	0.49	662	0.33	0.47	1782	0.40	0.49	744	0.35	0.48	1444
6	0.25	0.43	1693	0.26	0.44	1350	0.30	0.46	1436	0.28	0.45	1264	0.37	0.48	1133	0.29	0.45	1193	0.40	0.49	1446
7	0.40	0.49	1374	0.36	0.48	1380	0.35	0.48	1317	0.37	0.48	1265	0.38	0.48	1291	0.27	0.45	1510	0.40	0.49	1444
8	0.29	0.46	1604	0.36	0.48	1328	0.35	0.48	1425	0.28	0.45	1264	0.40	0.49	1271	0.33	0.47	1139	0.47	0.50	1444
9	0.29	0.46	1546	0.30	0.46	1355	0.32	0.47	1428	0.26	0.44	1265	0.42	0.49	1066	0.36	0.48	950	0.40	0.49	1539
10	0.32	0.47	1449	0.34	0.47	1354	0.32	0.47	1426	0.30	0.46	1263	0.29	0.45	1358	0.29	0.46	1170	0.39	0.49	1350

Note: For total individual income from benefits and other sources and for total household labour income, the share of missing cases is computed for positive cases only. IEMB and wave 1 are excluded.

Table 12: descriptive statistics of main income components, by imputation status and wave

Imputation status	mean	sd	min	max	p25	p50	p75	iqr	N	wave
Monthly gross earnings from first job (employees)										
Not imputed	1816.38	1355.83	0	8333	875.00	1506.74	2410.09	1535.09	22819	2
Imputed	1822.92	1252.92	39	8333	884.73	1585.94	2416.86	1532.13	1913	2
Not imputed	1849.05	1371.55	1	8333	886.97	1541.67	2500.00	1613.03	20668	3
Imputed	1859.84	1299.09	1	8333	910.07	1625.17	2499.88	1589.81	1673	3
Not imputed	1891.87	1401.51	0	8333	910.00	1596.00	2500.00	1590.00	19685	4
Imputed	1926.31	1335.51	1	8333	900.36	1697.31	2600.34	1699.98	1376	4
Not imputed	1930.41	1443.23	1	8333	930.00	1600.00	2505.55	1575.55	18797	5
Imputed	1921.77	1283.38	12	8333	933.21	1697.17	2599.99	1666.78	1378	5
Not imputed	1985.92	1466.26	0	8333	980.00	1650.00	2600.00	1620.00	16625	6
Imputed	1941.71	1405.15	13	8333	921.25	1599.95	2599.77	1678.52	1667	6
Not imputed	2025.77	1489.14	0	8333	1000.00	1680.49	2666.67	1666.67	16500	7
Imputed	1908.89	1337.90	1	8333	903.08	1646.47	2623.67	1720.59	1293	7
Not imputed	2083.64	1509.60	0	8333	1040.00	1733.33	2746.94	1706.94	15345	8
Imputed	2053.26	1474.31	1	8333	1029.83	1707.54	2747.00	1717.17	1583	8
Monthly gross earnings from first job (self employed)										
Not imputed	1619.03	2046.05	-15000	8333	370.08	1000.00	2110.89	1740.81	2230.00	2
Imputed	1615.02	1778.94	0	8333	499.23	1094.07	2022.43	1523.20	1376.00	2
Not imputed	1614.57	1995.77	-8000	8333	390.00	966.24	2093.00	1703.00	2135.00	3
Imputed	1735.41	1750.04	0	8333	588.41	1198.41	2176.92	1588.51	1280.00	3
Not imputed	1598.73	1993.17	-8750	8333	385.08	916.67	2083.33	1698.25	2099.00	4
Imputed	1805.69	1780.55	0	8333	594.86	1281.65	2251.39	1656.52	1180.00	4
Not imputed	1615.35	2008.84	-17889	8333	390.00	916.67	2093.00	1703.00	2016.00	5
Imputed	1639.53	1639.60	0	8333	588.46	1121.70	2165.50	1577.04	1176.00	5
Not imputed	1811.75	2179.60	-5000	8333	416.67	1000.00	2302.57	1885.90	1739.00	6
Imputed	1807.46	1889.57	0	8333	614.42	1227.83	2223.58	1609.16	1233.00	6
Not imputed	1695.48	2348.49	-41667	8333	390.00	964.81	2138.52	1748.52	1689.00	7
Imputed	2010.49	2108.81	0	8333	636.38	1272.91	2500.00	1863.62	1203.00	7
Not imputed	1727.26	2623.35	-54167	8333	400.00	964.81	2250.00	1850.00	1580.00	8
Imputed	1857.23	1944.89	0	8333	602.51	1299.99	2315.68	1713.17	1159.00	8
Monthly gross earnings from second job										
Not imputed	433.97	964.25	1	8333	80.00	200.00	400.00	320.00	2208.00	2
Imputed	470.51	775.09	1	6003	100.23	240.00	500.82	400.59	274.00	2
Not imputed	453.04	924.12	1	8333	90.00	200.00	450.00	360.00	1789.00	3
Imputed	617.77	1365.22	10	8333	116.72	266.80	500.00	383.28	254.00	3
Not imputed	444.10	955.81	1	8333	90.00	200.00	450.00	360.00	1914.00	4
Imputed	587.10	1153.60	5	8333	124.98	300.00	522.52	397.54	205.00	4
Not imputed	446.39	953.90	1	8333	85.00	200.00	420.00	335.00	1675.00	5
Imputed	566.54	1245.03	1	8333	100.35	250.23	439.34	338.99	168.00	5
Not imputed	462.73	949.40	1	8333	80.00	200.00	465.00	385.00	1812.00	6
Imputed	518.90	1035.45	5	8333	100.00	220.26	500.33	400.33	342.00	6
Not imputed	537.58	1139.28	1	8333	100.00	200.00	500.00	400.00	1483.00	7
Imputed	554.78	1193.44	5	8333	100.00	200.74	552.35	452.35	281.00	7
Not imputed	441.16	846.85	1	8333	100.00	200.00	496.00	396.00	1519.00	8
Imputed	480.42	781.01	1	8333	103.00	249.87	541.00	438.00	302.00	8
Annual income from interests and dividends										
Not imputed	1098.43	4741.06	1	100000	25.00	100.00	500.00	475.00	11292.00	2
Imputed	1166.50	4367.20	0	100000	50.02	199.77	603.04	553.02	4042.00	2
Not imputed	1205.90	5223.93	1	100000	30.00	100.00	500.00	470.00	9659.00	3
Imputed	1282.45	4850.99	0	100000	50.01	199.84	750.06	700.05	3889.00	3
Not imputed	1228.68	5033.00	1	100000	30.00	120.00	500.00	470.00	10256.00	4
Imputed	1347.04	4424.67	0	98002	50.07	200.14	803.55	753.48	4016.00	4
Not imputed	1273.17	5519.53	1	100000	30.00	130.00	500.00	470.00	9397.00	5
Imputed	1236.11	3990.26	0	100000	50.42	200.60	800.05	749.63	3246.00	5
Not imputed	1509.96	6157.12	1	100000	40.00	190.00	600.00	560.00	8933.00	6
Imputed	1209.52	4593.21	0	100000	50.36	200.16	694.75	644.39	4071.00	6
Not imputed	1573.12	6539.11	1	100000	38.00	150.00	600.00	562.00	8286.00	7
Imputed	1398.29	5194.71	0	100000	50.03	200.10	800.52	750.49	3207.00	7
Not imputed	1362.99	6116.94	1	100000	30.00	120.00	500.00	470.00	8472.00	8
Imputed	1150.22	5110.48	0	100000	41.55	150.75	510.81	469.26	5507.00	8

Note: For earnings from first job, the statistics are computed for applicable cases only. For earnings from second job and earnings from interests and dividends, the share of missing cases is computed for positive cases only. IEMB and wave 1 are excluded.

5. FINAL UKHLS INCOME VARIABLES

5.1.   OVERVIEW

This section summarises each of the UKHLS derived income variables. A range of derived income variables are available on the public release datasets. They may refer to different income components, they can be measured personal or household, net or gross, or with household level deductions.

Most analysis of standards of living, income dynamics and poverty and low income tend to use net or disposable household income after taxes and other deductions i.e., the income that people have available for consumption or saving. In deriving net incomes, UKHLS follows the approach used by the Department for Work Pensions (DWP) for their Households Below Average Income (HBAI) data sets.[12] The measures have been produced in close collaboration with staff from the DWP and they are now used as the data source for official UK statistics on income dynamics (DWP 2018).

In what follows, we have included data set variable names for ease of reference to the documentation and data. We refer to the variables on the "special license" datasets which, unlike the "end user" datasets, are not top-coded.

In the publically released files, none of the UKHLS income variables have been adjusted to account for price changes over time. In order to compare incomes collected in different months and years, each net household income value should be adjusted by a price index to a common time period (see section 3.7). The value of the price index can easily be matched to a household's month of interview (**w_intdatey**, **w_intdatem**) in each survey year, so that necessary adjustments to incomes can be performed.

This section proceeds as follows. First, in subsection 5.2, we present the UKHLS individual level income variables. In subsection 5.3, we present the household level income variables. In subsection 5.4, we then present household level variables for deductions (housing costs and council tax). Finally, to complete the section, we provide a table summarising the UKHLS derived income variables (table 13) and a table summarising derived housing costs variables (table 14).

---

[12] Subject to the limitations imposed by the UKHLS questionnaire. There are some deductions from individual income and some incomes sources which are not available to us from the questionnaire. The missing deductions are: maintenance and child support payments made; and parental contributions to students living away from home. The missing income sources are: income in kind from free school meals and free TV licences; and Healthy Start vouchers.

5.2. INDIVIDUAL INCOME VARIABLES

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

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

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

Component 1: Labour income (**w_fimnlabnet_dv**)

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

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)". This is assumed to be reported net of tax.

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

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

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

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

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

> This includes receipts reported in the income data file where ***w_ficode*** equals [2] "a pension from a previous employer", or [3] "a pension from a spouse's previous employer". This is assumed to be reported net of tax.

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

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

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

Less commonly, a researcher may be interested in the individual components of unearned income, such as individual benefit amounts. These are recorded in a separate data file (**w_income**). This file lists all reported unearned sources (as discussed in section 3.2). There may be multiple receipts of income from the same source in this file. For example, a respondent may have multiple pensions from a previous employer. Multiple receipts of a given income source are summed. These sums are recorded into the variable **w_frmnthimp_dv**. Therefore, for a given income source, **w_frmnthimp_dv** is equal to the total value of all receipts from that source for the first income receipt, it is set to zero for the subsequent receipts. The same income source may get reported by

more than one member of the household. To avoid double counting when constructing household income totals, the derived variable **w_frjtkeep_dv** identifies which one should be included in the household total. Note, the **w_income** file does not include individual income amounts for non-respondents in responding households.

5.3. HOUSEHOLD INCOME VARIABLES

This subsection summarises the household income derived variables. Household-level income measures are included in the household level data files, **_w_hhresp._**

In order to compare incomes for households of different size and composition, each net household income value should be adjusted by an equivalence scale (see section 3.7). The public release files contain values of the OECD-modified equivalence scale for each household (**w_ ieqmoecd_dv**). Equivalisation can be performed by dividing each household's income value by the equivalisation value provided.

At the household level, **_w_fihhmnnet1_dv_** is the net household monthly income. It is the sum of net monthly incomes from all household members (including proxies and within household non-respondents, see **_w_fimnnet_dv_**). It can be decomposed into the six subcomponents: net labour income (**w_fihhmnlabnet_dv**), miscellaneous income (**w_fihhmnmisc_dv),** private benefit income (**w_fihhmnprben_dv),** investment income (**w_fihhmninv_dv),** pension income (**w_fihhmnpen_dv),** and social benefit income (**w_fihhmnsben_dv**). The six subcomponents map to the six subcomponents of individual net income described in more detail in the previous subsection.

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

As indicated in the previous section, income components are imputed for all proxy and within household non-respondents. Hence in UKHLS household income estimates are available for all households – including where some household members non-respond. Users may decide to drop cases based on such imputed data but they would then need to adjust their results to take into account the consequent sample selection ie. their results would no longer be representative of the UK population.

5.4. HOUSEHOLD DEDUCTIONS

This subsection summarises the variables including household level deductions. The deductions cover housing costs and local taxes. Again, these household-level measures are included in the household level data files, **w_hhresp.**

HOUSING COSTS

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

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

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

For those paying mortgages, **w_xpmg_dv** is monthly total mortgage payments including imputation. The variable **w_xpmg_if** is the imputation flag for this variable. Most definitions of housing costs for

purposes of measuring income after housing costs seek to exclude repayments of capital included in mortgage payments and only include interest payments. **_w_xpmgint_dv_** is the estimated interest within **_w_xpmg_dv_**. For short period mortgages it is based on data on current interest rates times the outstanding principal and for mortgages with more than two years to run based on a standard repayment mortgage formula.

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

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

LOCAL TAXES

Council tax is a UK local tax that is often deducted from household income. In UKHLS, council tax liability is currently estimated for Great Britain, though not for Northern Ireland. Council tax liability (net council tax) is included in the variable **w_ficountax_dv**. It is equal to gross council tax applying any single person discount and then subtracting any council tax reduction. The variable **_w_fihhmnnet3_dv_** is equal to **_w_fihhmnnet1_dv_** less council tax liability and any council tax reduction. **_w_fihhmnnet4_dv_** is equal to **_w_fihhmnnet3_dv_** and adjusted for housing benefit reported in the household questionnaire (see above). For details of UK council tax, see section 3.2 starting "local taxes".

Table 13. Income components, income aggregates and their availability for type of respondent

						Available for		
Variable name	Label	Sum of	Data file	Imputation	Imputation flag	Individual respondents	Proxy respondents	Individual non respondents in responding households
Individual income variables								
1. payg_dv	gross pay per month in current job: last payment	Inapplicable	indresp	No	x	Yes	No	No
2. payn_dv	net pay per month in current job: last payment	Inapplicable	indresp	No	x	Yes	No	No
3. payu_dv	usual pay per month if different from last	Inapplicable	indresp	No	x	Yes	No	No
4. paygu_dv	usual gross pay per month: current job	Inapplicable	indresp	Yes	paygu_if	Yes	No	No
5. paynu_dv	usual net pay per month: current job	Inapplicable	indresp	Yes	paynu_if	Yes	No	No
6. seearngrs_dv	self employment earnings - gross	Inapplicable	indresp	Yes	seearngrs_if	Yes	No	No
7. seearnnet_dv	self employment earnings - net	Inapplicable	indresp	Indirectly	seearngrs_if	Yes	No	No
8. fiyrinvinc_dv	income from savings and investments, annual	Inapplicable	indresp	Yes	fiyrinvinc_if	Yes	Yes	No
9. j2pay_dv	pay in second job	Inapplicable	indresp	Yes	j2pay_if	Yes	No	No
10. j2paynet_dv	amount income component 1c: net earnings second job	Inapplicable	indresp	Indirectly	j2pay_if	Yes	Yes	No
11. frmnthimp_dv	Total income from benefit/other miscellaneous income source, including imputed	Inapplicable	income	Yes	frmnthimp_if	Yes	No	No
Individual income aggregates								
12. fimnlabnet_dv	amount income component 1: net labour income	5, 7, 10	indresp	Yes	x	Yes	Yes	No
13. fimnmisc_dv	amount income component 2: miscellaneous income	11 (ficode=24, 27, 38)	indresp	Yes	x	Yes	Yes	No
14. fimnprben_dv	amount income component 3: private benefit income	11 (ficode=25, 26, 35)	indresp	Yes	x	Yes	Yes	No
15. fimninvnet_dv	amount income component 5: investment income	11 (ficode=4, 28, 29), 8*	indresp	Yes	x	Yes	Yes	No
16. fimnpen_dv	amount income component 6: pension income	11 (ficode=2, 3)	indresp	Yes	x	Yes	Yes	No
17. fimnsben_dv	amount income component 7: social benefit income	11 (ficode=1, 5-16, 18-23, 30-34, 36, 37, 39-41)	indresp	Yes	x	Yes	Yes	No
18. fimnlabgrs_dv	total monthly labour income gross	4, 6, 9	indresp	Yes	fimnlabgrs_if	Yes	Yes	No
19. fibenothr_dv	Total income from benefits and other sources	13, 14, 15, 16, 17	indresp	Yes	fibenothr_if	Yes	No	No
20. fimngrs_dv	total monthly personal income gross	8*, 18, 19	indresp	Yes	fimngrs_if	Yes	Yes	No
21. fimnnet_dv	total net personal income - no deductions	8*, 12, 19	indresp	Indirectly	fimngrs_if	Yes	Yes	No
Household income aggregates								
22. fihhmninv_dv	total household investment income: month before interview	8* (sum over fhousehold members)	indresp	Yes	x	Yes	Yes	Yes
23. fihhmnlabnet_dv	total net household labour income: month before interview	12 (sum over fhousehold members)	indresp	Yes	x	Yes	Yes	Yes
24. fihhmnmisc_dv	total household miscellaneous income: month before interview	13 (sum over fhousehold members)	indresp	Yes	x	Yes	Yes	Yes
25. fihhmnprben_dv	total household private benefit income: month before interview	14 (sum over fhousehold members)	indresp	Yes	x	Yes	Yes	Yes
26. fihhmnpen_dv	total household pension income: month before interview	15 (sum over fhousehold members)	indresp	Yes	x	Yes	Yes	Yes
27. fihhmnsben_dv	total household social benefit income: month before interview	17 (sum over fhousehold members)	indresp	Yes	x	Yes	Yes	Yes
28. fihhmnlabgrs_dv	total gross household labour income: month before interview	18 (sum over fhousehold members)	indresp	Yes	x	Yes	Yes	Yes
29. fihhmngrs_dv	gross household income: month before interview	19 (sum over fhousehold members)	indresp	Yes	fihhmngrs_if	Yes	Yes	Yes
30. fihhmnnet1_dv	total household net income - no deductions	21 (sum over fhousehold members)	indresp	Indirectly	fihhmngrs_if	Yes	Yes	Yes

Note: * indicates the annual investment income has been divided by 12 to derive a monthly equivalent. Ficode in column 3 refers to the unique value assigned to each unearned income source. The unearned income sources are corresponding ficodes are listed in table 15.

Table 14. Council tax and housing costs variables

Variable name	Description
	Council tax variables
ficountax_dv	gross council tax applying any single person discount and then subtracting any council tax reduction.
fihhmnnet3_dv	household net income less council tax liability and any council tax reduction (fihhmnnet1_dv-ficountax_dv)
	Housing costs variables
rentgrs_dv	monthly gross rent including any housing benefit received (household questionnaire only)
rent_dv	the monthly rent net of housing benefit (equal to rentgrs_dv where no housing benefit is received). Where a household reports 100% housing benefit (in which case no rent value is reported), the value is set equal to housing benefit reported in the individual questionnaire or imputed if missing.
hbadjust_dv	housing benefit adjustment equal to housing benefit reported in the household questionnaire minus the sum of housing benefit reported in the individual questionnaires.
fihhmngrs1_dv	gross household income adjusted for housing benefit reported in the household questionnaire (fihhmngrs_dv + hbadjust_dv)
fihhmnnet4_dv	fihhmnnet3_dv adjusted for housing benefit reported in the household questionnaire (fihhmnnet3_dv + hbadjust_dv)

6. COMPARISONS TO A SPECIALIST INCOME SURVEY

6.1. OVERVIEW

Currently there is no longitudinal counterpart with which to validate the UKHLS household net income series.[13] Instead, we compare UKHLS estimates of the income distribution to those from a cross-sectional income survey. This is a useful quality check as longitudinal measures of change, such as income mobility and poverty transitions, are essentially formed from the difference between two cross-sectional estimates. Much of the comparisons that follow replicate those reported in Jenkins (2011) for the BHPS.[14]

6.2. COMPARISON WITH THE HOUSEHOLDS BELOW AVERAGE INCOME SERIES

As the UKHLS net income series aims to replicate the Households Below Average Income (HBAI) series, our cross-sectional counterpart is the HBAI. The HBAI is the data source for official UK statistics on the income distribution. It is based on a specialist income survey (the Family Resources Survey (FRS)) that undergoes extensive editing and imputation by the UK Department for Work and Pensions, and benefits from their access to administrative records and knowledge of the tax and benefit system. The HBAI is considered to be of high quality. In what follows, we use the terms HBAI and FRS interchangeably.

The key differences between the surveys are that i) the HBAI fieldwork period corresponds to a financial year rather than the two calendar year fieldwork period in UKHLS and ii) new immigrants are covered by HBAI but not UKHLS. To make the coverage of the two surveys comparable, we remove new immigrant households since 2010 from the HBAI data and also the IEMB subsample from UKHLS. To make the reference periods comparable, we construct UKHLS financial year samples by combining data from multiple waves ie. months 4-15 from wave n are combined with months 16-24 from wave n-1 and months 1-3 from wave n+1. We include all currently available UKHLS data in our main analysis but note we must exclude data collected before April 2010 (wave one) and after March 2015 (wave eight) where it is not possible to construct full financial year samples. Our main validations therefore correspond to the financial years 2010-15.[15]

---

[13] UKHLS respondents were asked to give consent to linked to administrative income sources. This will allow for some longitudinal validations in the future (see section 3.5).

[14] Jenkins (2011) shows that the BHPS household net income series closely mirrors official poverty statistics. While Francesconi, Sutherland and Zantomio (2011) find that the BHPS individual earnings series diverges from the Family Resources Survey over time.

[15] Minor differences in definition across the two data sources remain. See footnote 12.

All figures are expressed in 2015-16 prices using the bespoke monthly CPI price index produced by the Office for National Statistics (see section 3.7). Household net income is equivalised using the OECD-modified scale[16] and the HBAI amounts converted to a monthly equivalent. All figures are weighted to the UK population using the relevant cross-sectional weights. Note, our construction of financial year subsamples in UKHLS requires pooling weights across waves and thus we perform an additional rescaling of the UKHLS weights so that later waves are not under-represented.

Estimates of selected quantiles of the income distribution are presented in figure 5. Reassuringly, the estimates from the two surveys line up closely and show a similar time series pattern. However, we do see systematic differences, although they are very small in magnitude. UKHLS, relative to HBAI, tends to overestimate percentiles 1, 5, and 10 by a small amount. In contrast, HBAI gives marginally higher estimates for the 95th and 99th percentiles. For individuals in the very richest households (p99), the gap tends to be bigger. The latter likely reflects the known difficulties in measuring the incomes of the very rich.[17]

We also compare estimates of inequality. Here, we trim the top and bottom 1 percent of each sample as measures of inequality maybe sensitive to outliers in the data. Figures 6-14 plot trends in percentile ratios (90-10, 90-50, 50-10 and 75-25), the standard UK poverty rate[18] and inequality indices (Gini, Theil, Atkinson, and GE2) for each data source. Looking across the measures, we see that measured inequality is very similar in both data sets, although typically higher according to the HBAI. This reflects the differences at the bottom half of the distribution as above. Note, the comparisons in Jenkins (2011) document a similar inequality difference between the HBAI and BHPS at its later waves.

The comparisons so far suggest that the UKHLS measure of net income fares well at the population level. We now compare each survey at the level of population subgroups. Figure 15 examines the share of the population that falls into each of eight family types (pensioner couple, single female (male) pensioner, couple with (without) children, single with children, single female (male) without

---

[16] We follow the standard international convention of giving a scale rate of one to a single-person household. This differs from what is sometimes done in the UK. See the discussion on "Equivalisation and Price Indices" (section 3.7).

[17] The HBAI includes an adjustment to a very small number of 'very rich' individuals using information from personal income tax returns data. So in principle, it should provide better estimates of those with the very highest incomes. See Burkhauser et al. (2018).

[18] The share of individuals living in households with income 60% below the median.

Figure 5. Selected quantiles of household net income
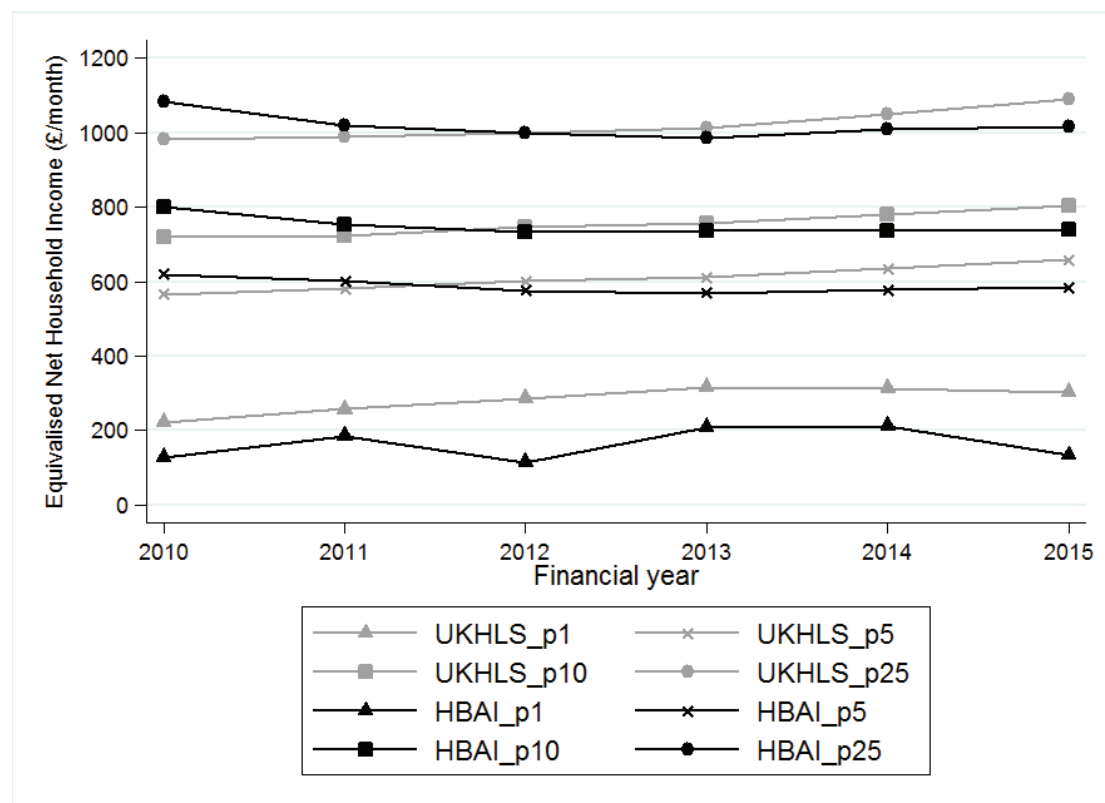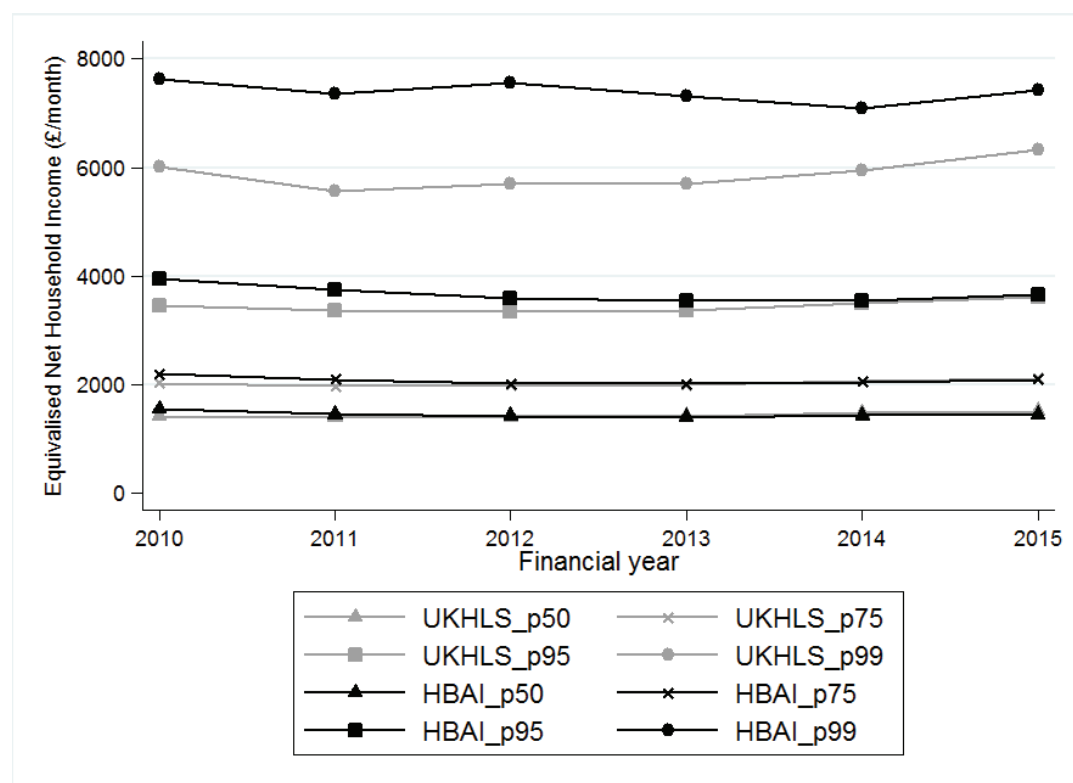
children). The series for the full population line up nicely but there are some differences. The family type breakdowns suggest that UKHLS over-estimates the proportion of pensioner couples at the population level (e.g., 15.1% vs 13.5% in 2015) and conversely underestimates the share of couples without children (e.g., 16.2% vs. 19.0% in 2015). Figures 16-20 replicate the family breakdowns but by income quintile. For example, for the bottom quintile (figure 16), UKHLS over-estimates the share of pensioners in particular single female pensioners (eg. 10.1 % vs. 7.6 % in 2015). Overestimation of single female pensioners in the bottom quintile was also found by Jenkins (2011) for the BHPS. The converse is that couples with children and without children are slightly under-represented in the bottom quintile (eg. 32.3 vs. 35.1 and 7.4 vs. 10.6 in 2015, respectively).

We are also able to decompose household income into five subcomponents (figures 21-26: earnings, state benefits, occupation pensions, investments and other income). For this analysis, we work with (unequivalised) gross household income, as its subcomponents are directly comparable across the data sets. Figures 21 and 22 show that the means of income and earnings are close in both surveys. For example, for 2015 we see a mean total monthly income of £4021 in UKHLS and £3930 in HBAI, just over a two percent difference. The corresponding figures for earnings are £2929 and £2935, a 0.2 percent difference. However, UKHLS captures more state benefit income relative to HBAI (£625 vs. £562 in 2015 or a 10 percent difference). Therefore, the two surveys are very similar in terms of earnings, but UKHLS overestimates the share of state benefit income relative to HBAI (figure 27). Differences across the other sources, as a proportion of household income, are typically small. Figures 28-32 present the decompositions for each quintile separately and the broad pattern seen at the population level is seen in each quintile.

In summary, this section has compared various estimates of the income distribution from UKHLS and the HBAI. The comparisons represent two independent attempts at measuring the UK income distribution. Estimates from each data source have given a remarkably similar picture of living standards and how they have changed over the period of our comparisons.

We have seen similar levels and trends for various income percentiles, inequality measures, the composition of household incomes across the distribution, and the types of benefit unit we see across the distribution. Where there are differences, we see that UKHLS, relative to HBAI, tends to slightly overestimate income at the bottom of the distribution and underestimate it slightly at the top. This means that overall UKHLS tends to have slightly lower inequality levels. UKHLS has proportionally more pensioners and household incomes contain a slightly larger share of state benefit income.

While UKHLS mirrors HBAI in terms of cross-sectional estimates of poverty, it is not possible to construct longitudinal income measures in HBAI, so we cannot directly validate the UKHLS measures in this regard. However, the fact that the two data sources give very similar cross-sectional estimates, leads us to conclude that UKHLS forms a sound basis for longitudinal research on the income distribution.

A caveat to the above should, however, be noted. The HBAI is not a gold-standard benchmark but rather also based on a survey. It suffers from non-response (56 percent at the latest survey), like UKHLS, which weighting seeks to address. As such, our validations reflect two independent attempts to estimate the same population parameters. We interpret the fact that UKHLS tracks this leading cross-section income survey as indicative that it starts out and remains of high quality over time.
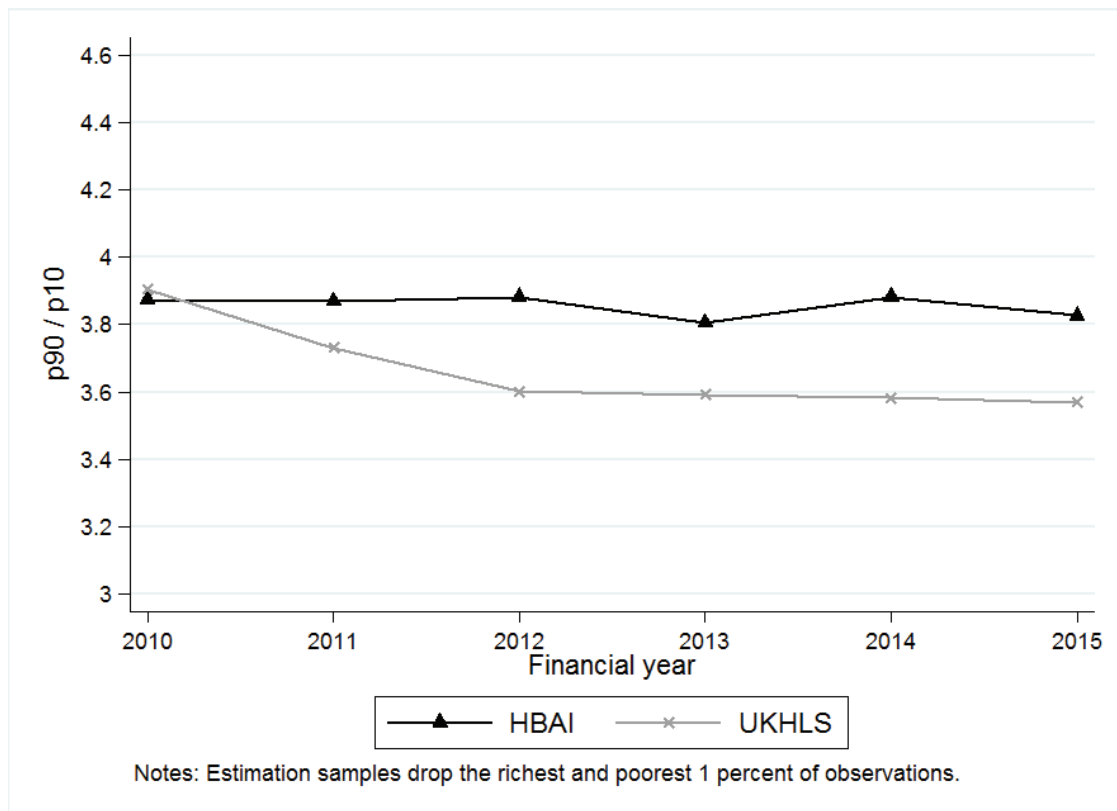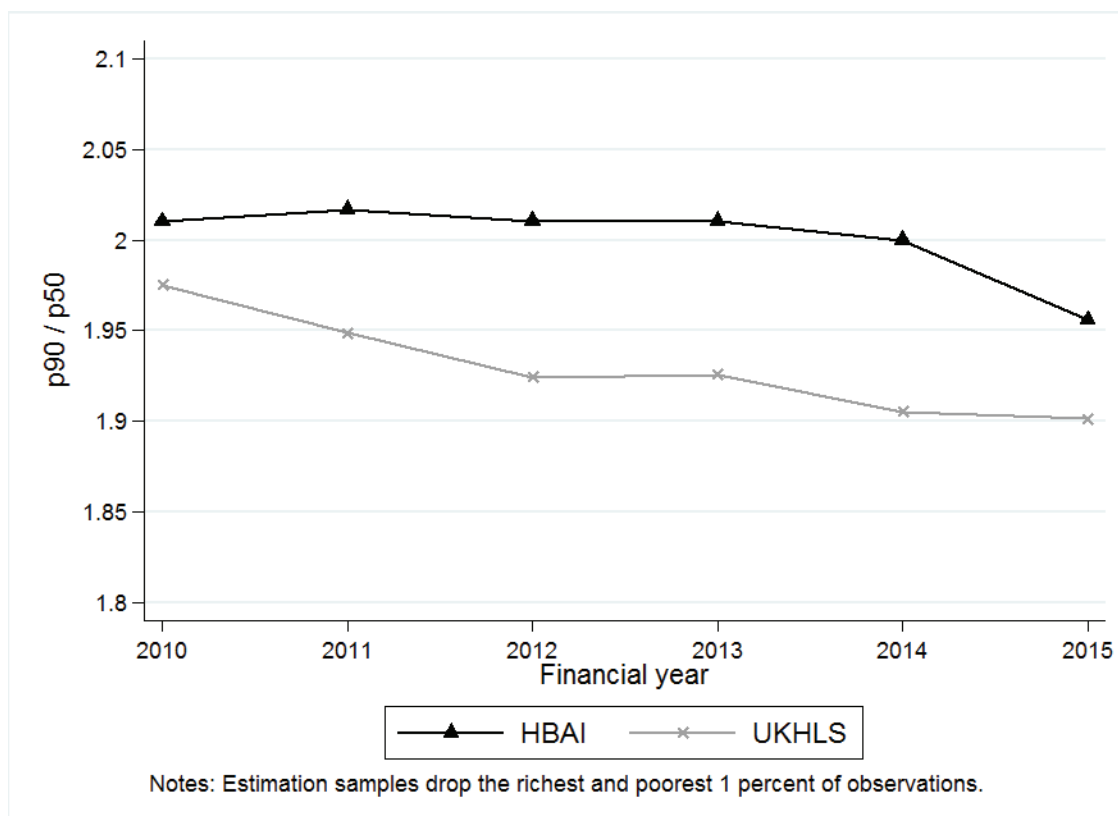
Figure 6. P90/P10

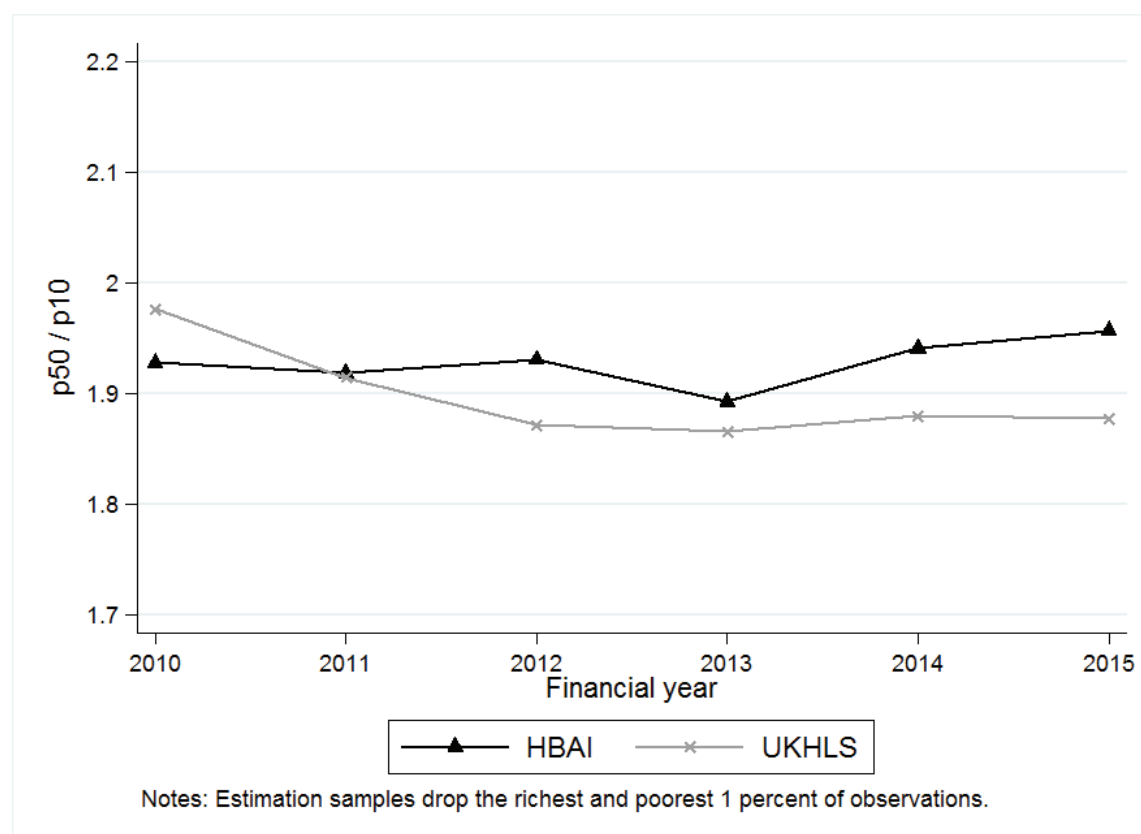Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 7. P90/P50

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 8. P50/P10

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 9. P75/P25

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
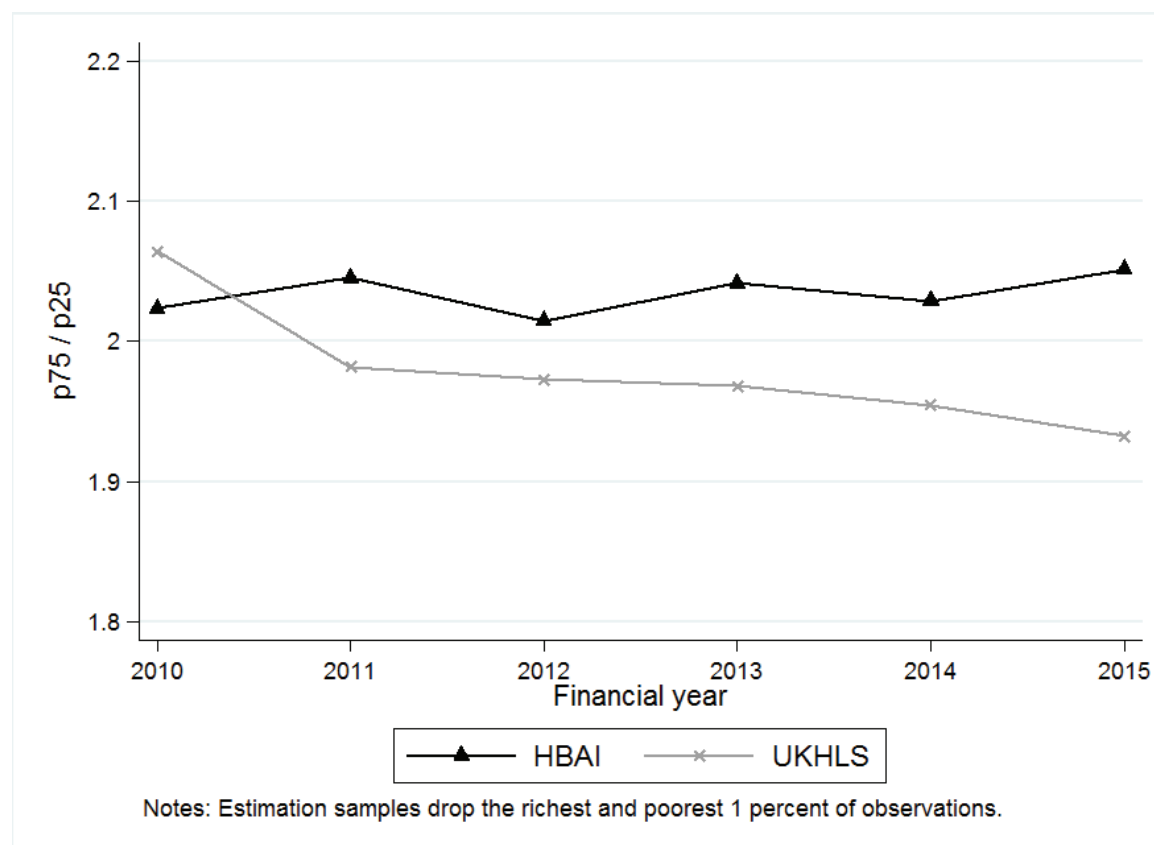
Figure 10. Share with household income 60% below median

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 11. Gini

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

57

Figure 12. Theil

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 13. GE2

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 14. Atkinson

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
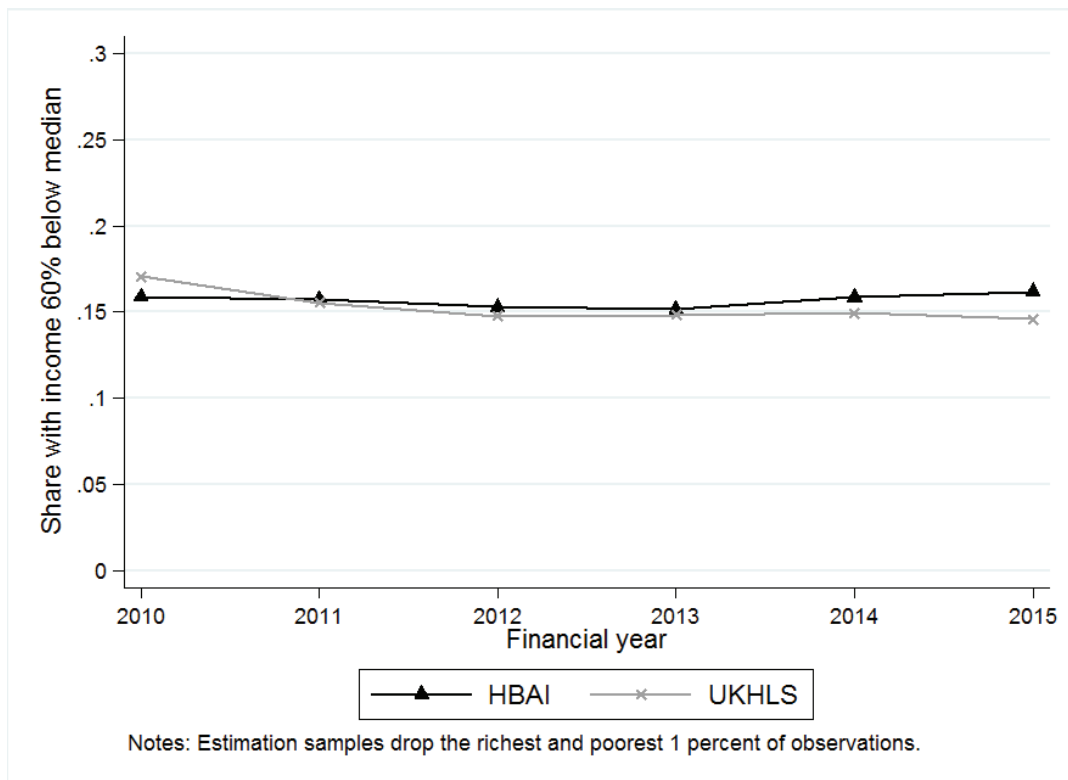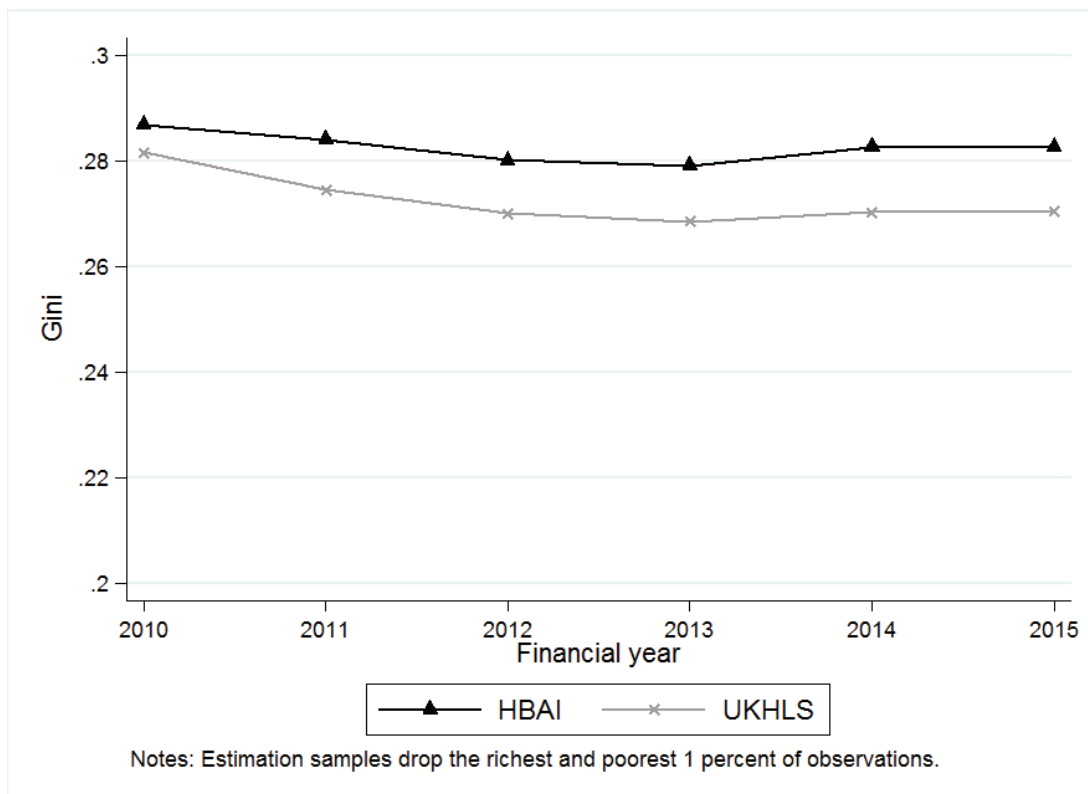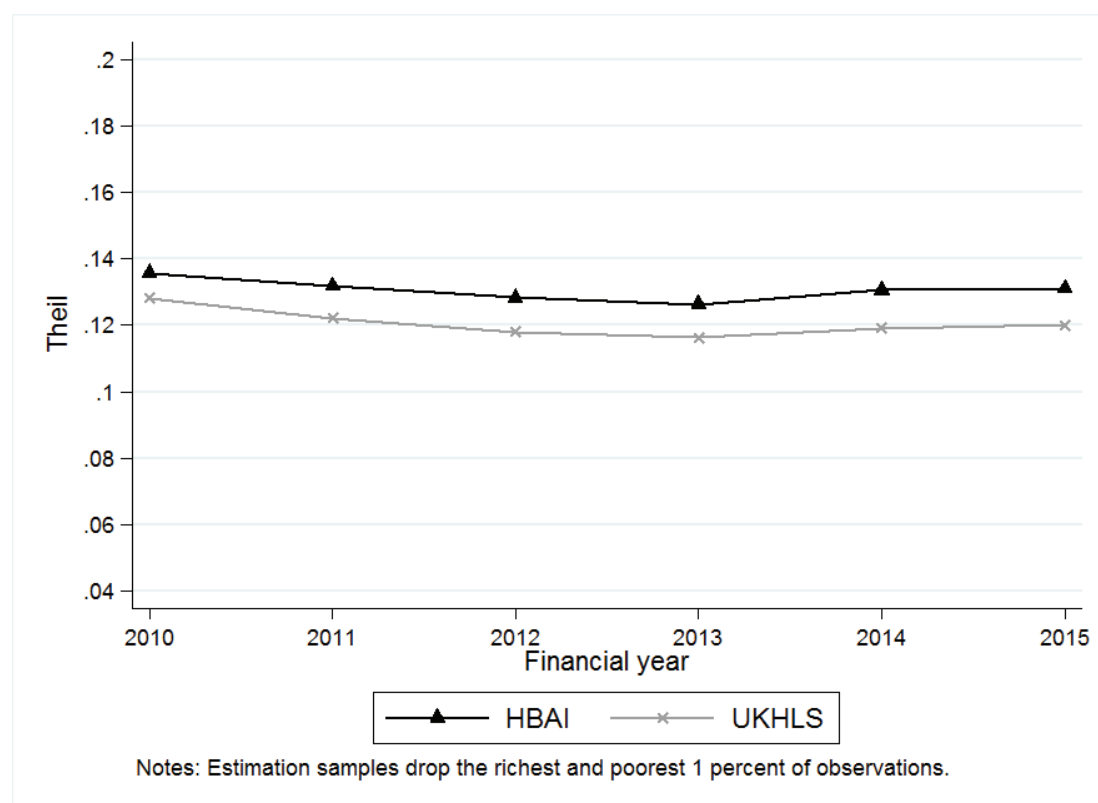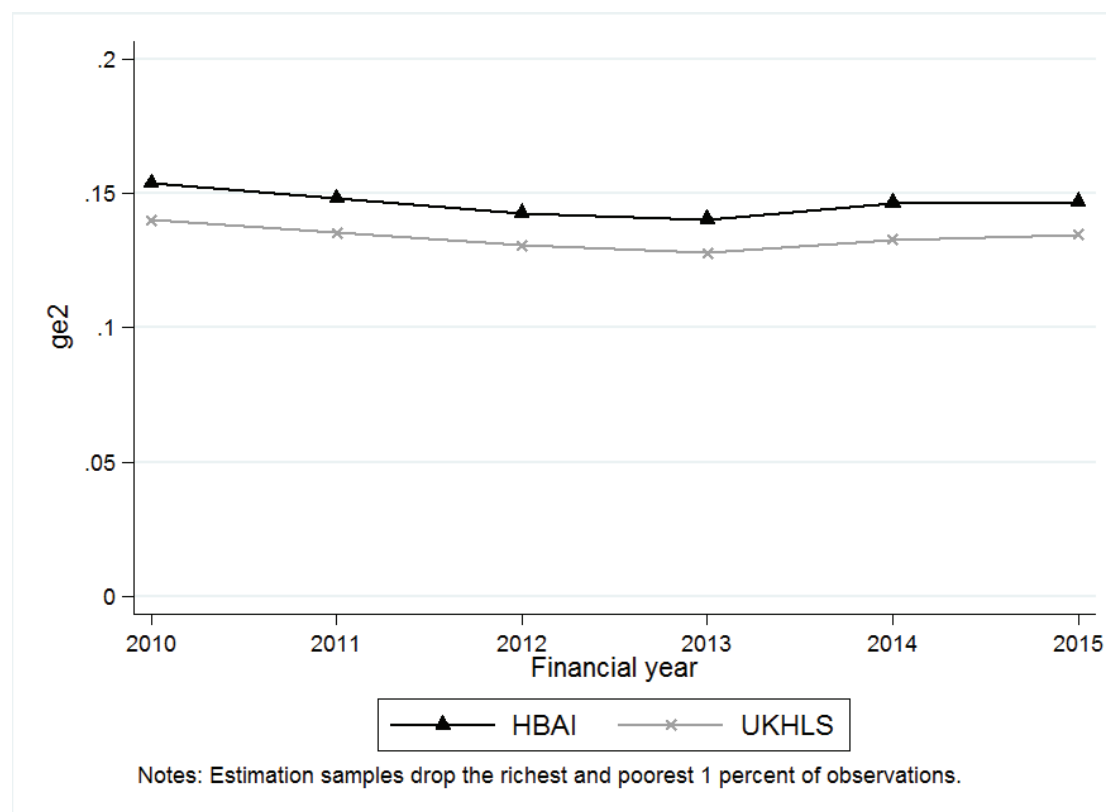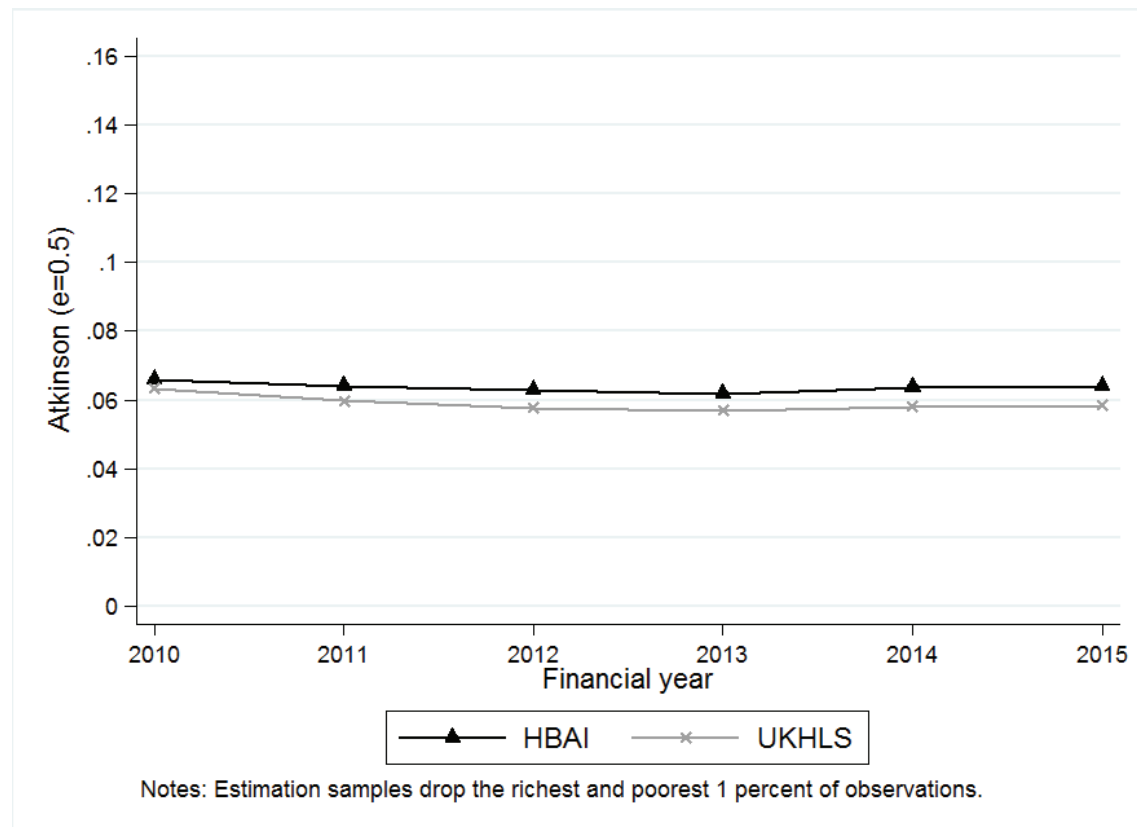
Figure 15. Benefit unit composition

Figure 16. Benefit unit composition: lowest quintile

Benefit unit composition: lowest quintile by financial year
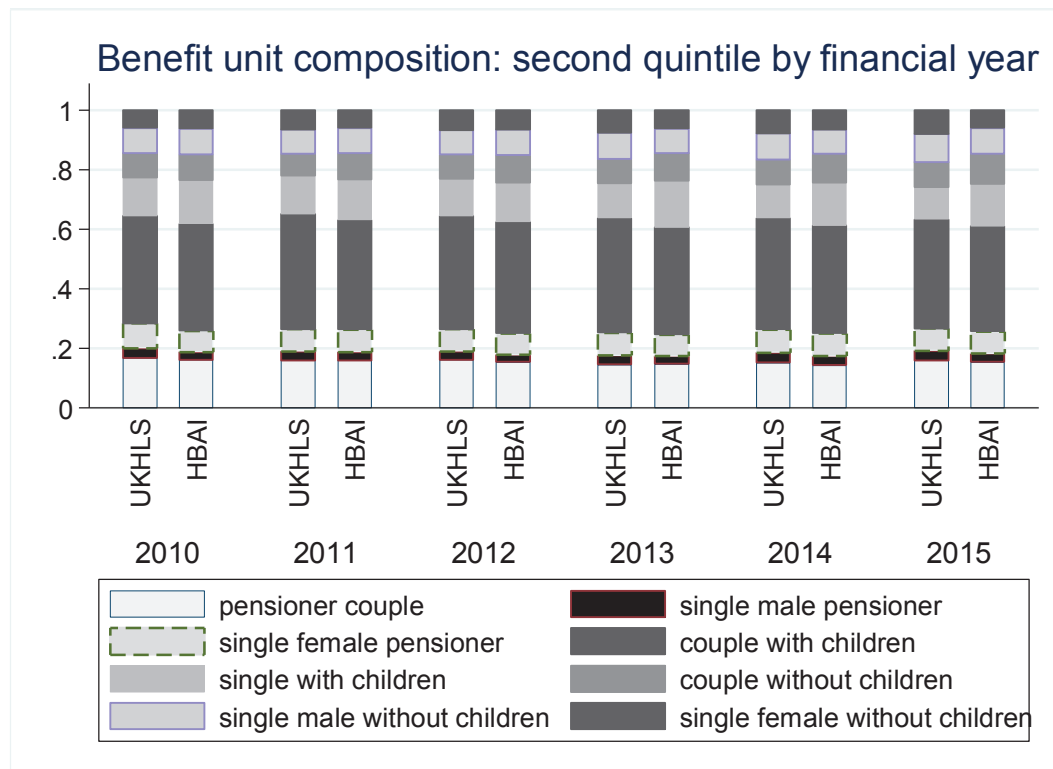
Legend:
- pensioner couple
- single female pensioner
- single with children
- single male without children
- single male pensioner
- couple with children
- couple without children
- single female without children
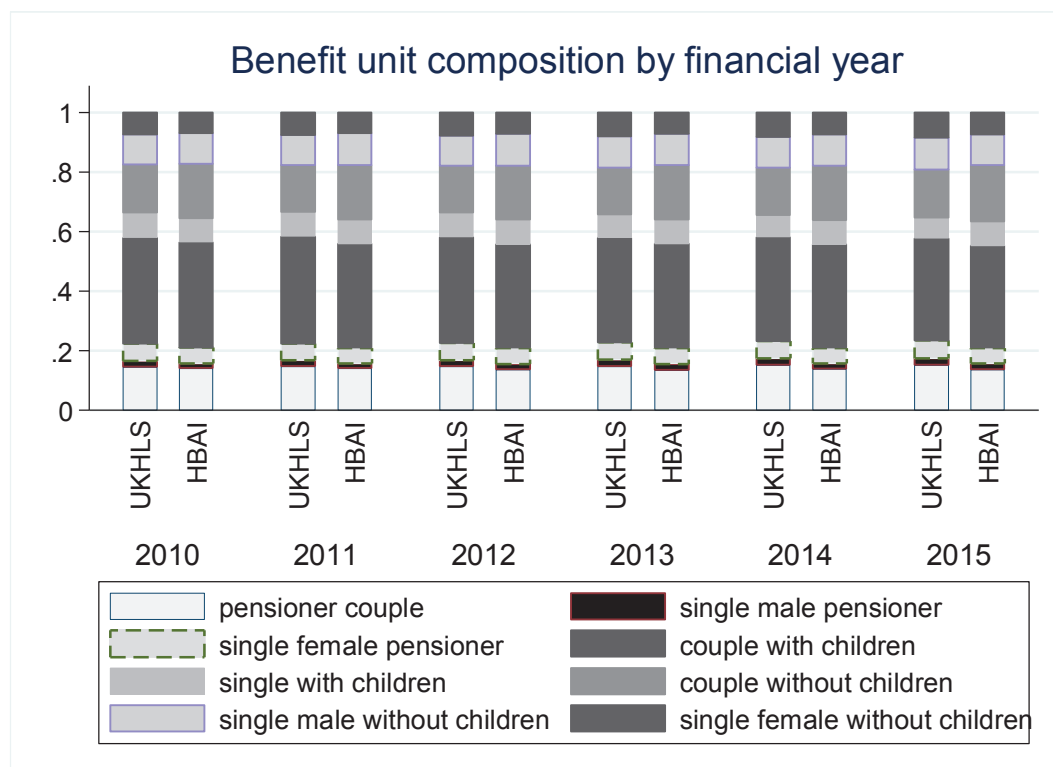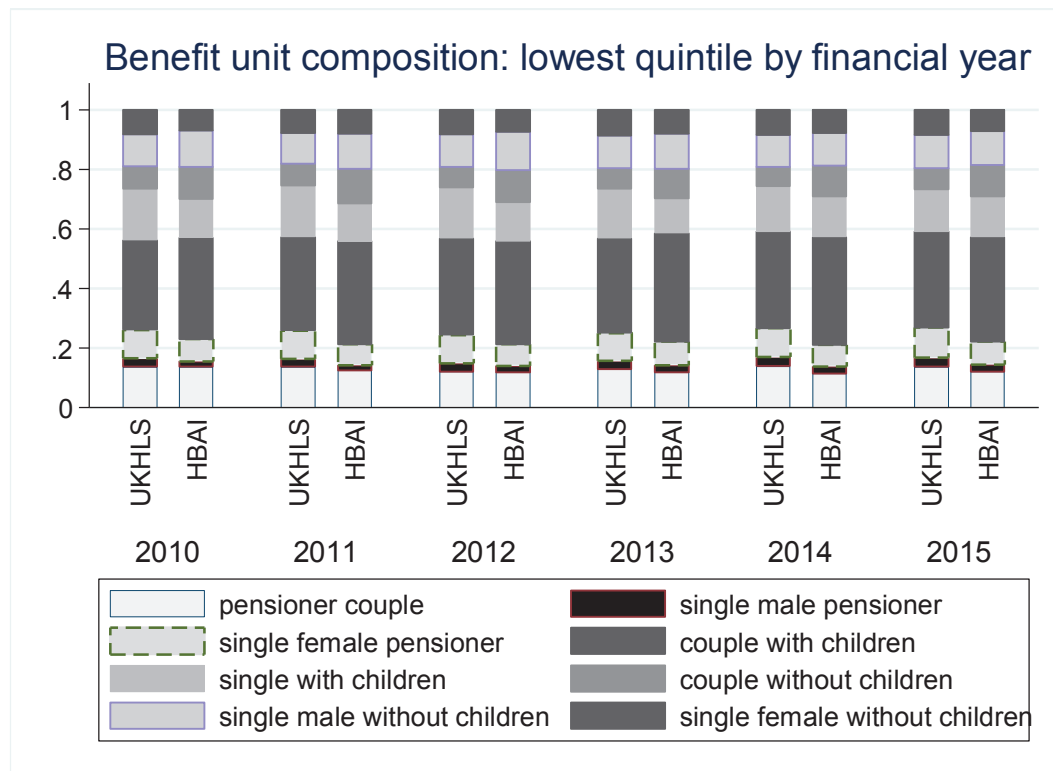
Figure 17. Benefit unit composition: second quintile

Benefit unit composition: second quintile by financial year

Legend:
- pensioner couple
- single female pensioner
- single with children
- single male without children
- single male pensioner
- couple with children
- couple without children
- single female without children

Figure 18. Benefit unit composition: third quintile

Benefit unit composition: third quintile by financial year

Legend:
- pensioner couple
- single female pensioner
- single with children
- single male without children
- single male pensioner
- couple with children
- couple without children
- single female without children
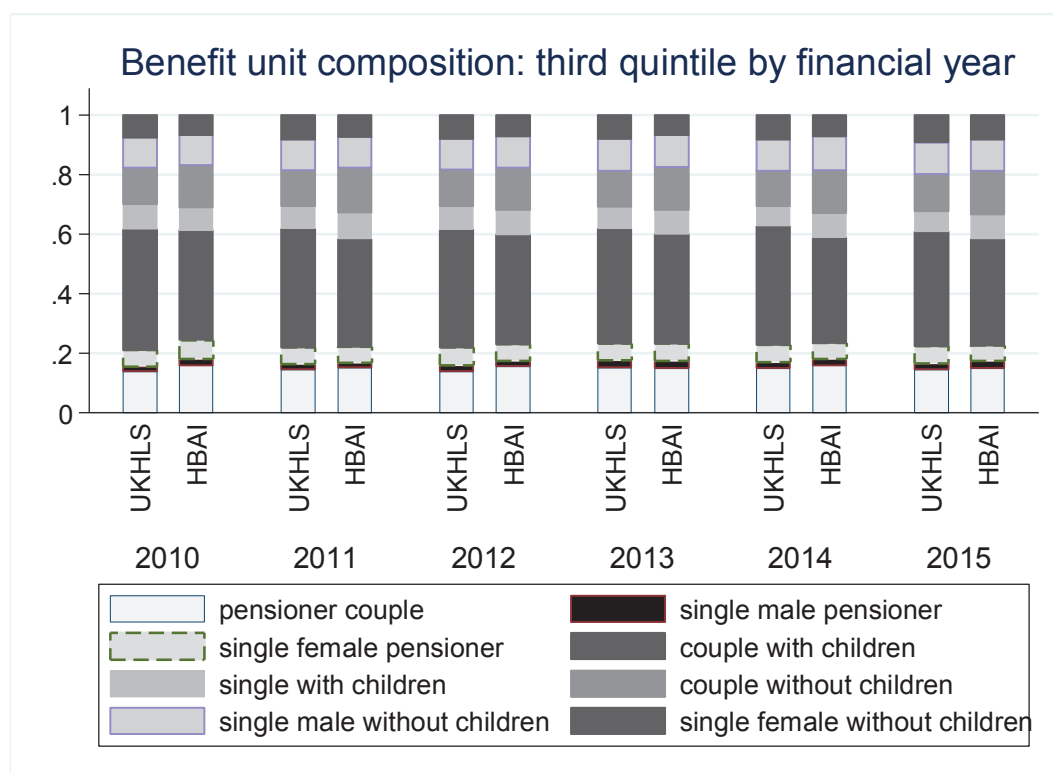
Figure 19. Benefit unit composition: fourth quintile

Benefit unit composition: fourth quintile by financial year

Legend:
- pensioner couple
- single female pensioner
- single with children
- single male without children
- single male pensioner
- couple with children
- couple without children
- single female without children
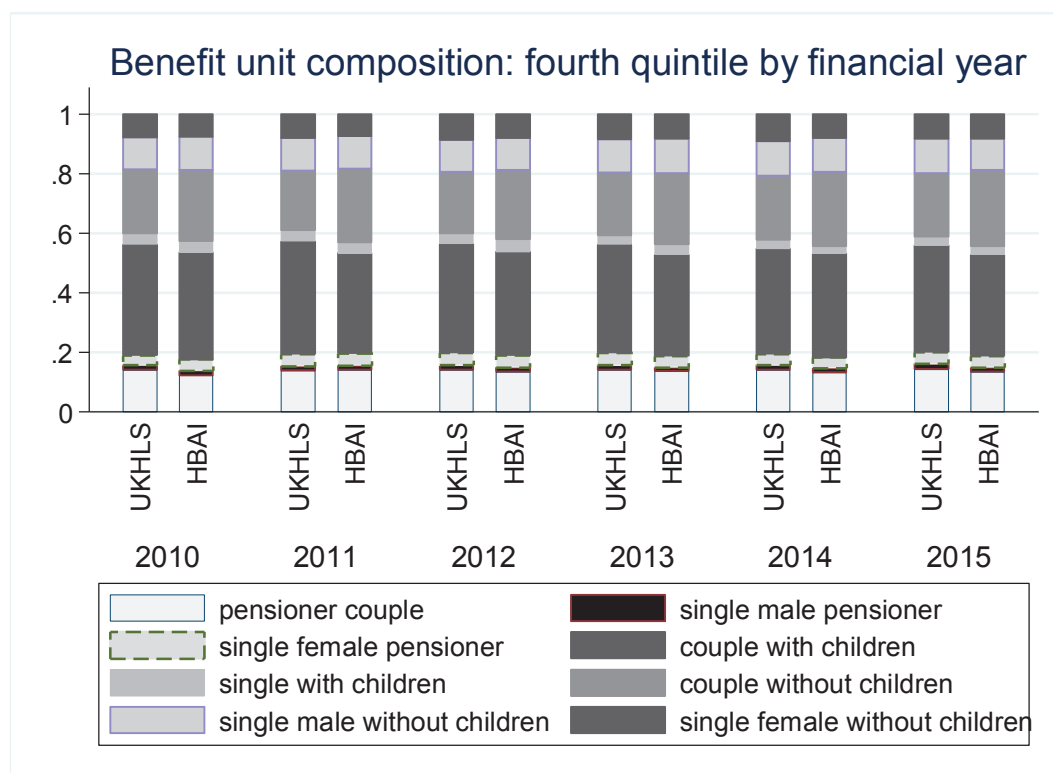
Figure 20. Benefit unit composition: top quintile

Benefit unit composition: top quintile by financial year

Legend:
- pensioner couple
- single female pensioner
- single with children
- single male without children
- single male pensioner
- couple with children
- couple without children
- single female without children

Figure 21. Mean gross household income

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 22. Mean gross household earnings

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 23. Mean state benefit income

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 24. Mean occupation pension income

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 25. Mean of "other" income

Notes: Estimation samples drop the richest and poorest 1 percent of observations.

Figure 26. Mean investment income

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
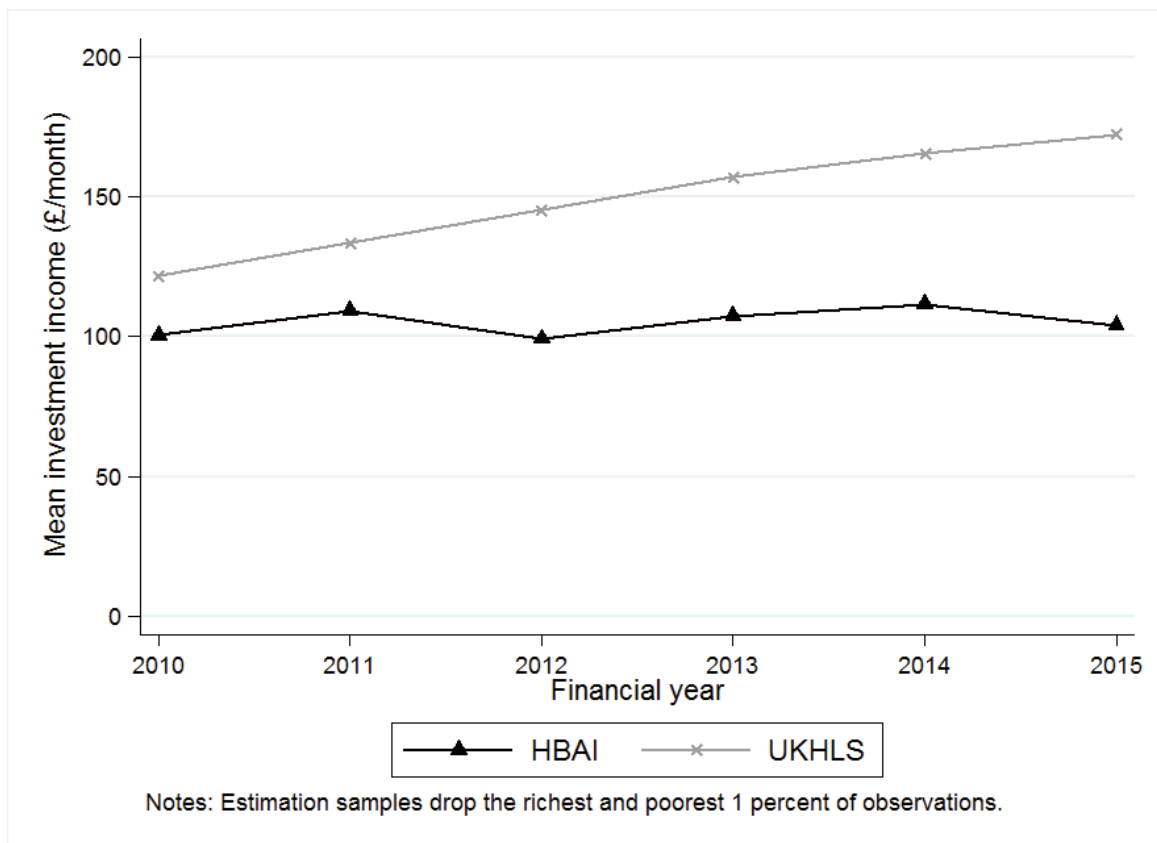
Figure 27. Decomposition of gross household income

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
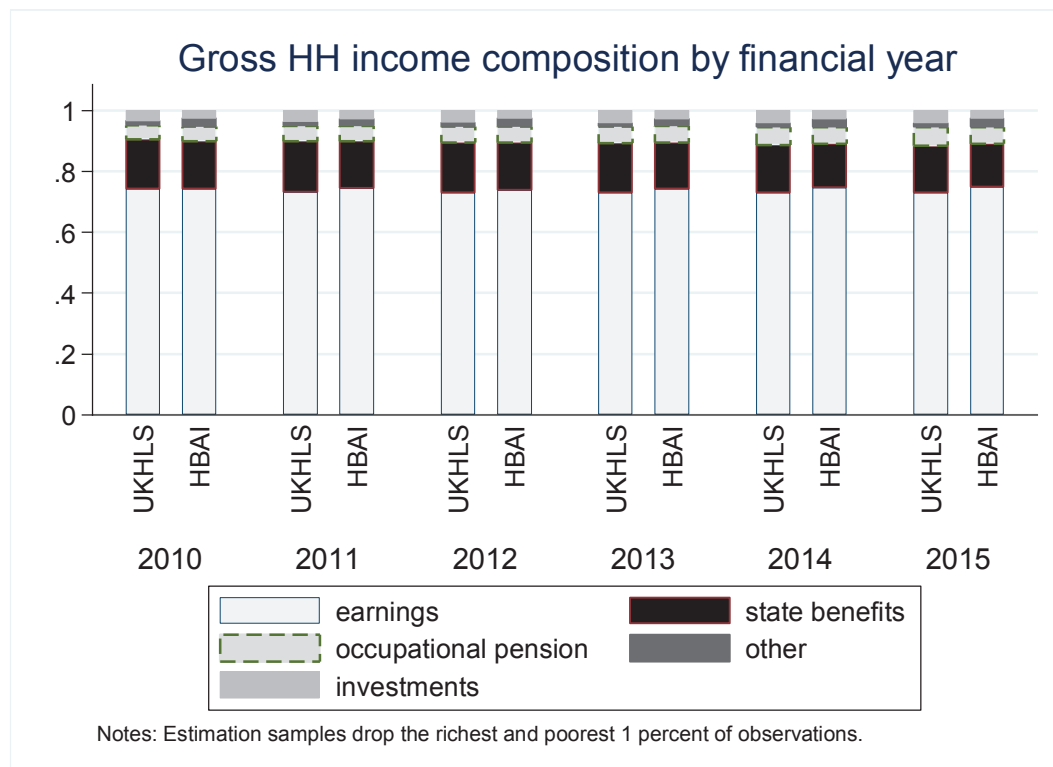
Figure 28. Decomposition of gross household income: lowest quintile

Gross HH income composition: lowest quintile by financial year

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
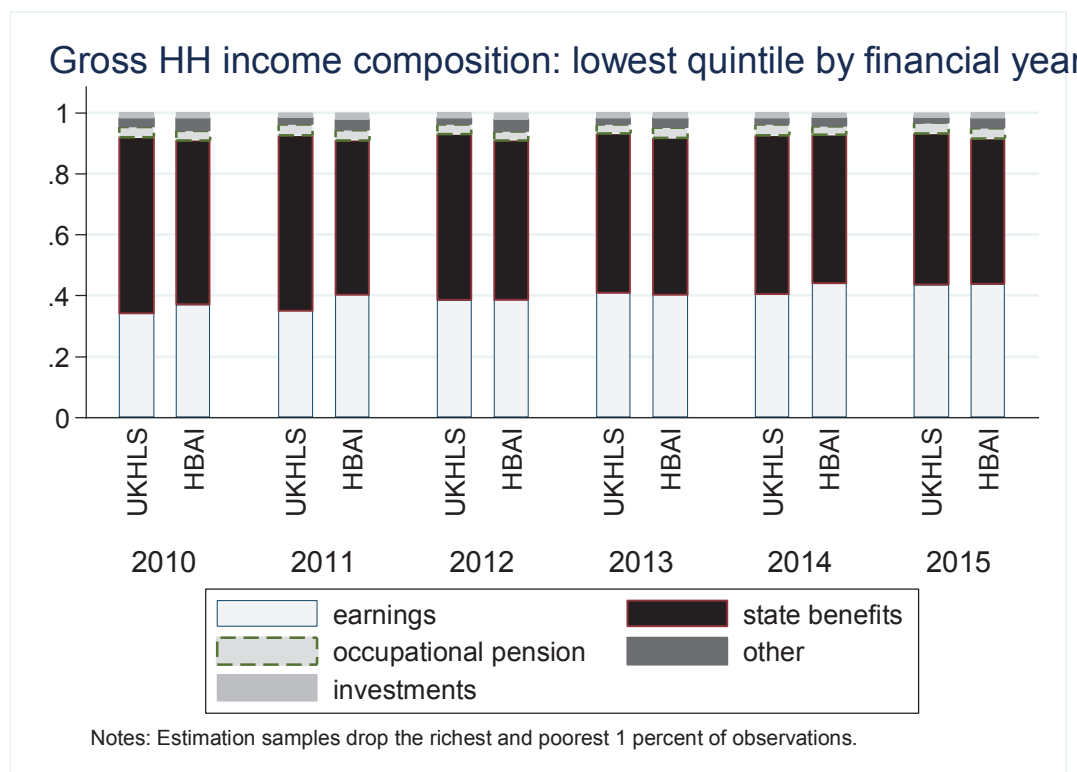
Figure 29. Decomposition of gross household income: second quintile

Gross HH income composition: second quintile by financial year

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
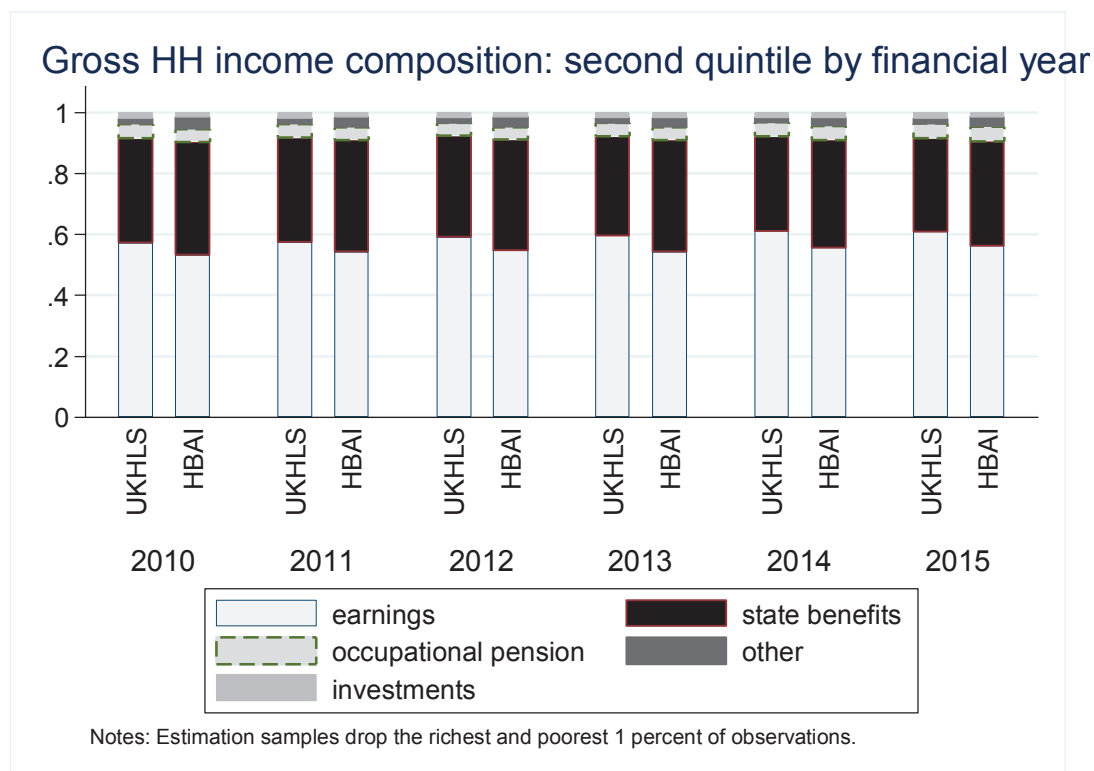
Figure 30. Decomposition of gross household income: third quintile

Gross HH income composition: third quintile by financial year

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
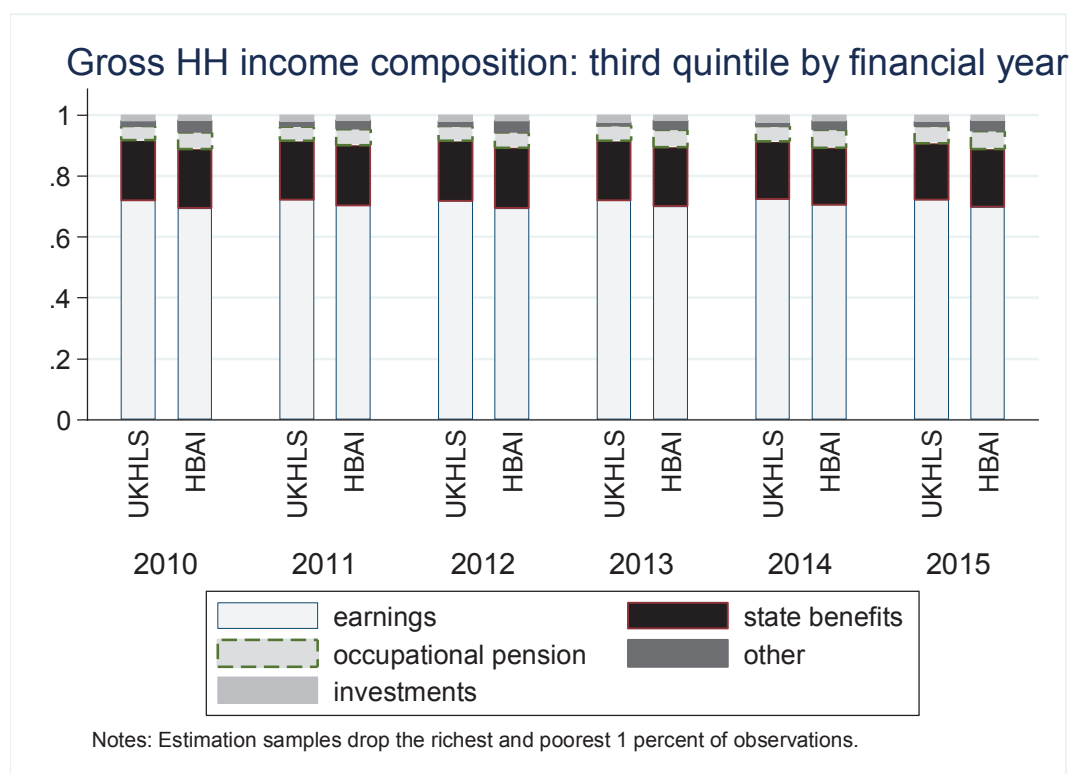
Figure 31. Decomposition of gross household income: fourth quintile

Gross HH income composition: fourth quintile by financial year

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
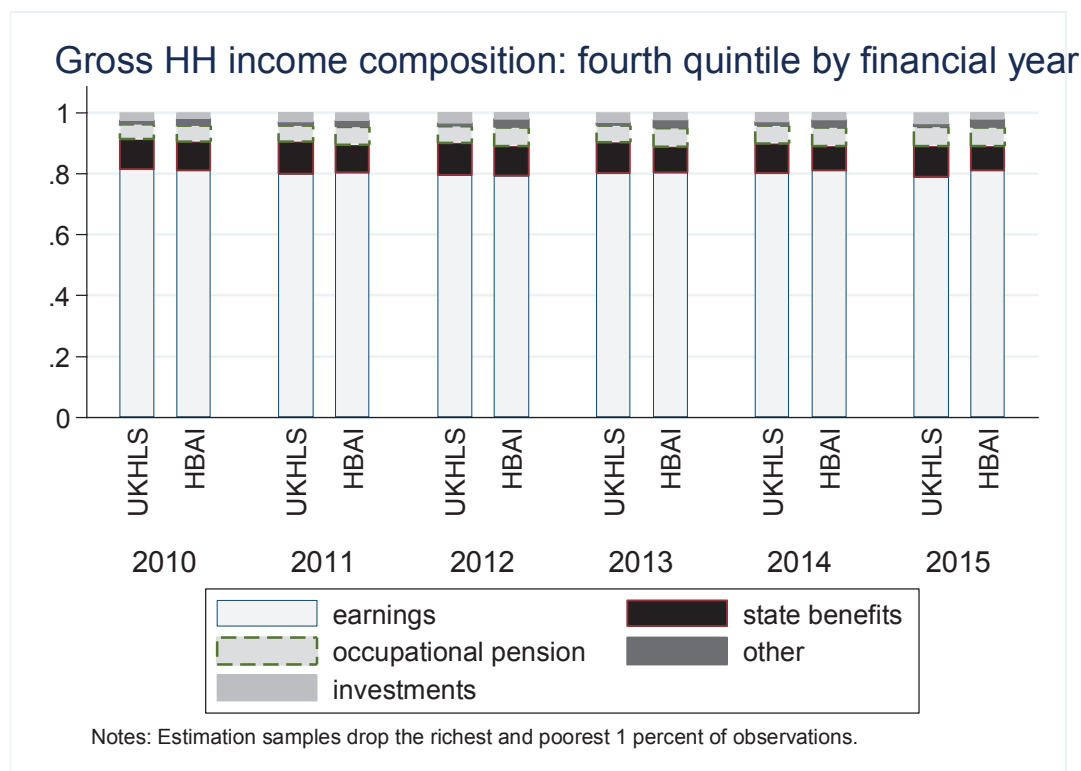
Figure 32. Decomposition of gross household income: top quintile

Gross HH income composition: top quintile by financial year

Legend:
- earnings
- occupational pension
- investments
- state benefits
- other

Notes: Estimation samples drop the richest and poorest 1 percent of observations.
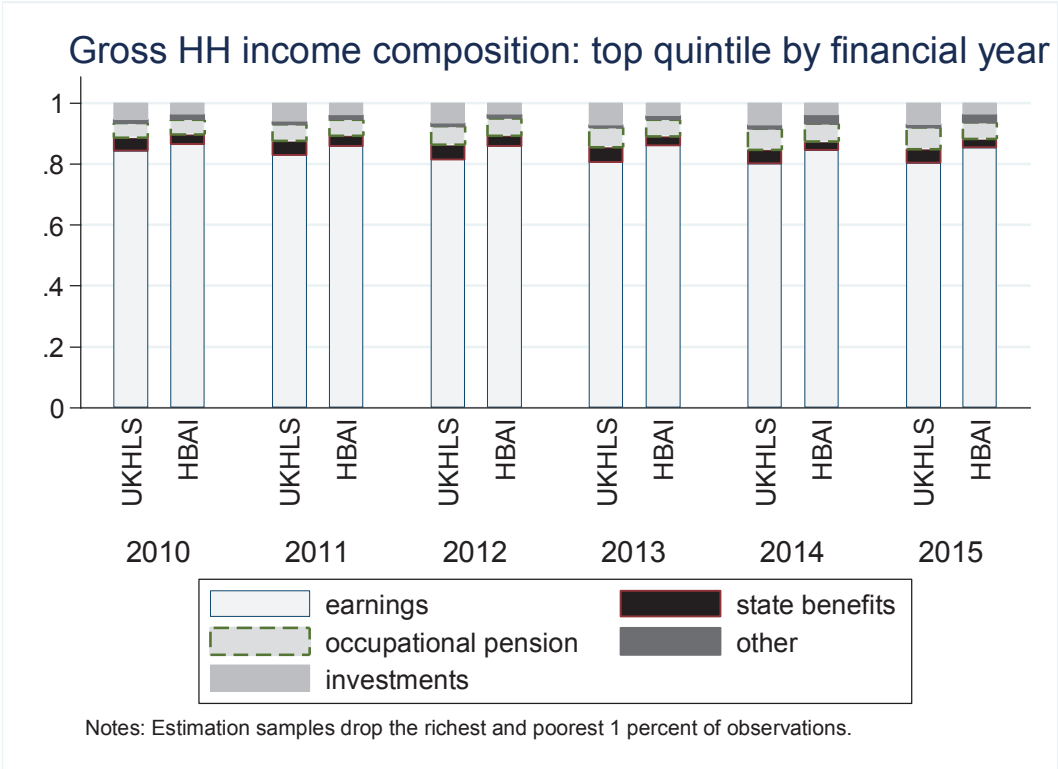
7. ADDITIONAL MATERIALS

7.1. ORIGINAL, TEMPORARY AND PERMANENT SAMPLE MEMBERS

Original sample members: i) individuals found at selected households in wave one, including individuals temporarily absent but otherwise resident; ii) individuals in households eligible for the inclusion in EMBS or IEMBS, and from a target ethnic minority; iii) children born to a female OSM and living with the OSM mother at the wave following their birth; iv) individuals from the BHPS sample component who were enumerated at the first wave of the BHPS sample they belong to (1991 BHPS sample, Scotland, Wales, or Northern Ireland boost samples); v) children born to a BHPS OSM (irrespective of the gender) before the inclusion of BHPS in UKHLS.[19] We attempt to retain OSM respondents as part of the sample as long as they live in the UK.

Temporary Sample Members: i) Individuals joining the household of an OSM after the sample selection/first interview, such as partners, house sharers or family members joining the household; ii) individuals in households eligible for the inclusion in EMBS, but not from a target ethnic minority; iii) individuals in households eligible for the inclusion in IEMBS, but not from a target ethnic minority or immigration background; iv) children born to male OSM and a female TSM and co-resident with the father (or any other OSM) at the survey wave following the child's birth. We attempt to interview TSM participants in successive waves as long as they live in the household of an OSM.

Permanent Sample members (PSM): TSM who do no longer live with OSM, but they are still followed as they can bring to the panel crucial information about the OSMs. At the moment, the only PSM are TSM who are fathers of an OSM child and are observed to live with this child the wave after their birth.

7.2. JOINT RECEIPT OF UNEARNED INCOME

In what follows, we have included data set variable names for ease of reference to the documentation and data. The variables referred to can be found on the file "w_income.dta", which lists all reported unearned sources.

1. In order to identify double counting, the individual questionnaire asks, for each unearned income source reported, whether a respondent receives the amount jointly and if so, who

---

[19] After the incorporation of BHPS into Understanding Society, OSM sample status is assigned in line with UKHLS following rules also to respondents from the BHPS sample component.

with (**w_frjt**). For example, at wave 7, 16.6 percent of unearned income sources were reported jointly. The 5 most common sources reported jointly were child benefit, child tax credit, rent from a property, council tax benefit and housing benefit.

2. We derive an indicator variable which identifies whether a reported unearned amount should be included when calculating household totals (**jtkeep_dv**). It takes the value 1 if the amount should be included or 0 if it should be ignored because it is counted elsewhere. The person number identified as the joint recipient is included in the variable **frjtpnmod_dv**.

3. As respondents sometimes report inconsistent information, the derivation of **jtkeep_dv** requires us to make some assumptions on whether a source should be counted once or twice in household totals. These are as follows:

Partner 1 report	Partner 2 report	Assumption
sole	didn't report the source	Include partner 1 amount in the household total
sole	sole	Include both partner amounts in the household total
joint	joint	Treat as one payment and include only one amount in the household total
joint	sole	Treat as one payment and include only one amount in the household total
joint	didn't report the source	Include partner 1 full amount in the household total

70

5. It can also be the case that a payment is identified as joint, but the two joint recipients have report different amounts. In this case, we need to make an assumption about which amount to keep in the total ie. which source will have **jtkeep_dv**=1 and which one **jtkeep_dv**=0. The assumption we make is that the larger of the two reports is the correct one and so include it in the household total. For each reported source, the assumption made is summarised in the variable **frjtmod_dv**.

6. If a non-respondent is listed as a joint recipient of a source, the imputed amount for the non-respondent is set to zero and the full amount of the responding partner is included in the total. Note, that imputed values for non-respondents are not included on the public release files. See section 4.

7. If a non-household member is listed as a joint recipient, the full amount reported is included in the household total.

If a respondent reports multiple payments of the same source, in the final derived amount for that source (**frmnthimp_dv**), we sum across all payments with **jtkeep_dv**=1. So for a given income source, **w_frmnthimp_dv** is equal to the total value of all receipts from that source for the first income receipt, it is set to zero for the subsequent receipts.

Additional tables and figures
Table 15: 41 unearned income sources collected in UKHLS

Pensions:
1)   ni retirement/state retirement (old age) pension
2)   pension, previous employer
3)   pension from a spouse's previous employer
4)   private pension/annuity
5)   widow's or war widow's pension

Social security benefits:
6)   widowed mother's allowance / widowed parent's allowance / bereavement allowance
7)   pension credit (incl. guarantee credit & saving credit)
8)   severe disablement allowance
9)   industrial injury disablement allowance
10)   disability living allowance
11)   attendance allowance
12)   carer's allowance (was invalid care allowance)
13)   war disablement pension
14)   incapacity benefit
15)   income support
16)   job seeker's allowance
17)   national insurance credits
18)   child benefit (incl. lone-parent child benefit payments)
19)   child tax credit
20)   working tax credit (incl. disabled person's tax credit)
21)   maternity allowance
22)   housing benefit
23)   council tax benefit
24)   foster allowance / guardian allowance
25)   rent rebate
26)   rate rebate
27)   employment and support allowance
28)   return to work credit
29)   in-work credit for lone parents

Other unearned income:
30)   educational grant (not student loan or tuition fee loan)
31)   trade union / friendly society payment
32)   maintenance or alimony
33)   payments from a family member not living here
34)   rent from boarders or lodgers (not family members) living here
35)   rent from any other property
36)   sickness and accident insurance

Other:
37)   other disability related benefit or payment
38)   any other regular payment
39)   income from any other state benefit

Added at wave 4:

40) universal credit

41) personal independence payment

Figure 33: construction of usual gross pay (w_paygu_dv)

	valid amount		yes							
Gross pay:		**Is it usual?**								
w_paygl		*w_payusl*								
w_paygwc			No		valid amount			Gross		
				Usual pay:		**Gross or net?:**			Finish	
				w_payu		*w_payug*				
				w_payuwc				Net	apply net gross	
									conversion and	
missing		missing		missing					finish	
								missing		
									impute	

Figure 34: construction of usual net pay (w_paynu_dv)

	valid amount		yes							
Gross pay:		**Is it usual?**								
w_paynl		*w_payusl*								
w_paynwc			No		valid amount			Net		
				Usual pay:		**Gross or net?:**			Finish	
				w_payu		*w_payug*				
				w_payuwc				Gross	apply gross net	
									conversion and	
missing		missing		missing		missing			finish	
									impute	

73

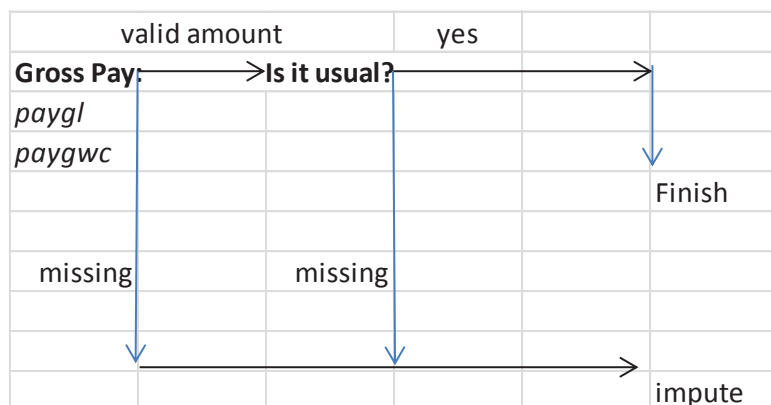**Figure 35: construction of gross pay last payment (w_payg_dv)**

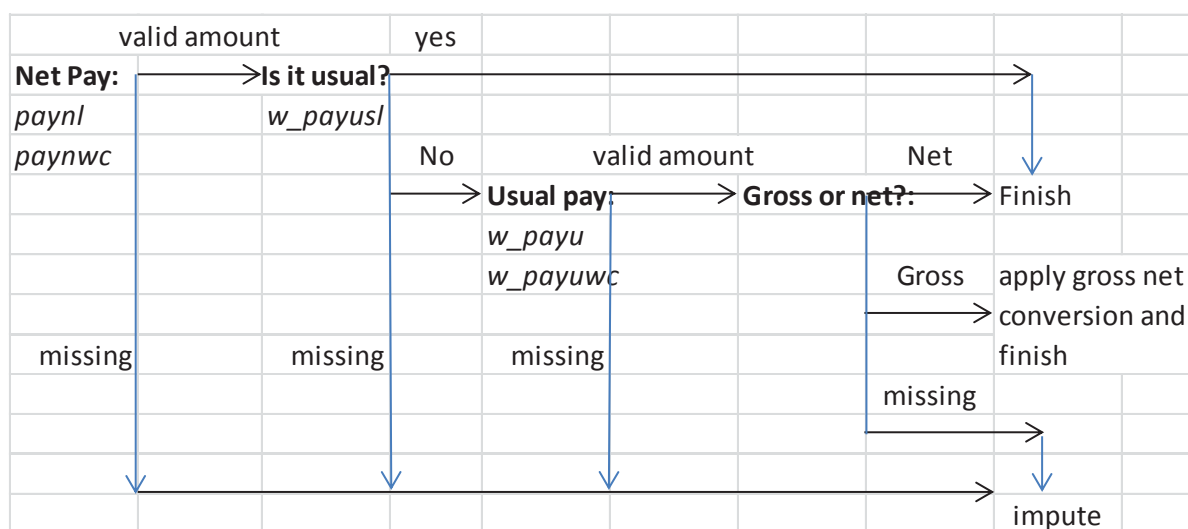**Figure 36: construction of net pay last payment (w_payn_dv)**

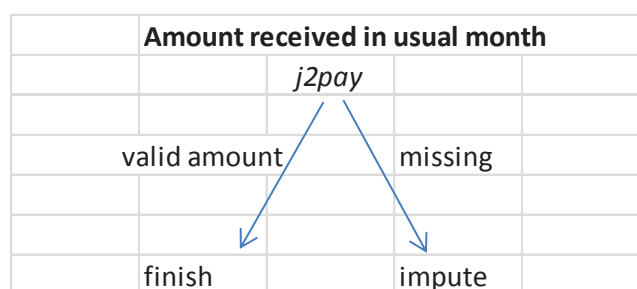**Figure 37: construction of gross second job income (w_j2pay_dv)**

Figure 38: construction of gross self-employment earnings (w_seearngrs_dv)

	valid amount	profit	gross	
	figure on accounts:	**profit or loss?:**	**Gross or net?:**	finish
		w_jsprls	*w_jsprtx*	
	w_jsprbm		*w_jsprni*	net
	w_jsprby4			apply net gross conversion and finish
	w_jsprem			
	w_jsprey4			missing
yes	*w_jsprf*			
		loss		finish
accounts prepared for inland revenue:	missing	missing		
w_jsaccs				impute
no/not yet				
	valid amount	gross		
	average monthly income:	**Gross or net?:**		finish
		w_jspytx		
		w_jspyni	net	
	w_jspayu			apply net gross conversion and finish
	w_jspayw			
	missing	missing		

Figure 39: construction of net self-employment earnings (w_seearnnet_dv)

	valid amount	profit	net	
	figure on accounts:	**profit or loss?:**	**Gross or net?:**	finish
		w_jsprls	*w_jsprtx*	
	w_jsprbm		*w_jsprni*	gross
	w_jsprby4			apply gross net conversion and finish
	w_jsprem			
	w_jsprey4			missing
yes	*w_jsprf*			
		loss		finish
accounts prepared for inland	missing	missing		
w_jsaccs				impute
no/not yet				
	valid amount	net		
	average monthly income:	**Gross or net?:**		finish
		w_jspytx		
		w_jspyni	gross	
	w_jspayu			apply gross net conversion
	w_jspayw			
	missing	missing		

75

Figure 40: construction of unearned income sources (w_frmnthimp_dv)

	Last amount received (sum multiple payments from the same source)					
	frval					
	frwc					
valid amount		missing				
finish		impute				
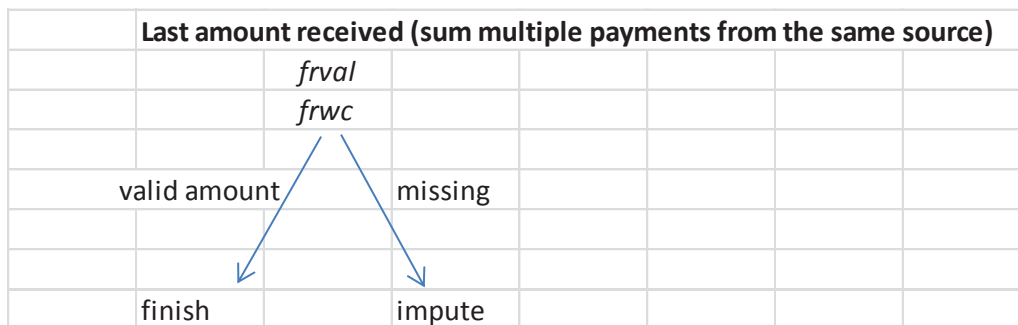
Figure 41: construction of income from investments and savings (w_fiyrinvinc_dv)

						valid amount				
			amount in						Finish	
			interest /							
			dividends:							
			w_fiyrdia							
		don't								
		know								
missing					No		valid/missing			
			500 or more			**100 or more**				
			w_fiyrdb1			*w_fiyrdb6*				
		yes								
missing					No					
			1000 or more							
			w_fiyrdb2							
		yes								
missing					No		valid/missing			
			5000 or more			**2500 or more**				
			w_fiyrdb3			*w_fiyrdb4*				
		yes								
missing					No					
			10,000 or more							
			w_fiyrdb5							
		yes								
impute										
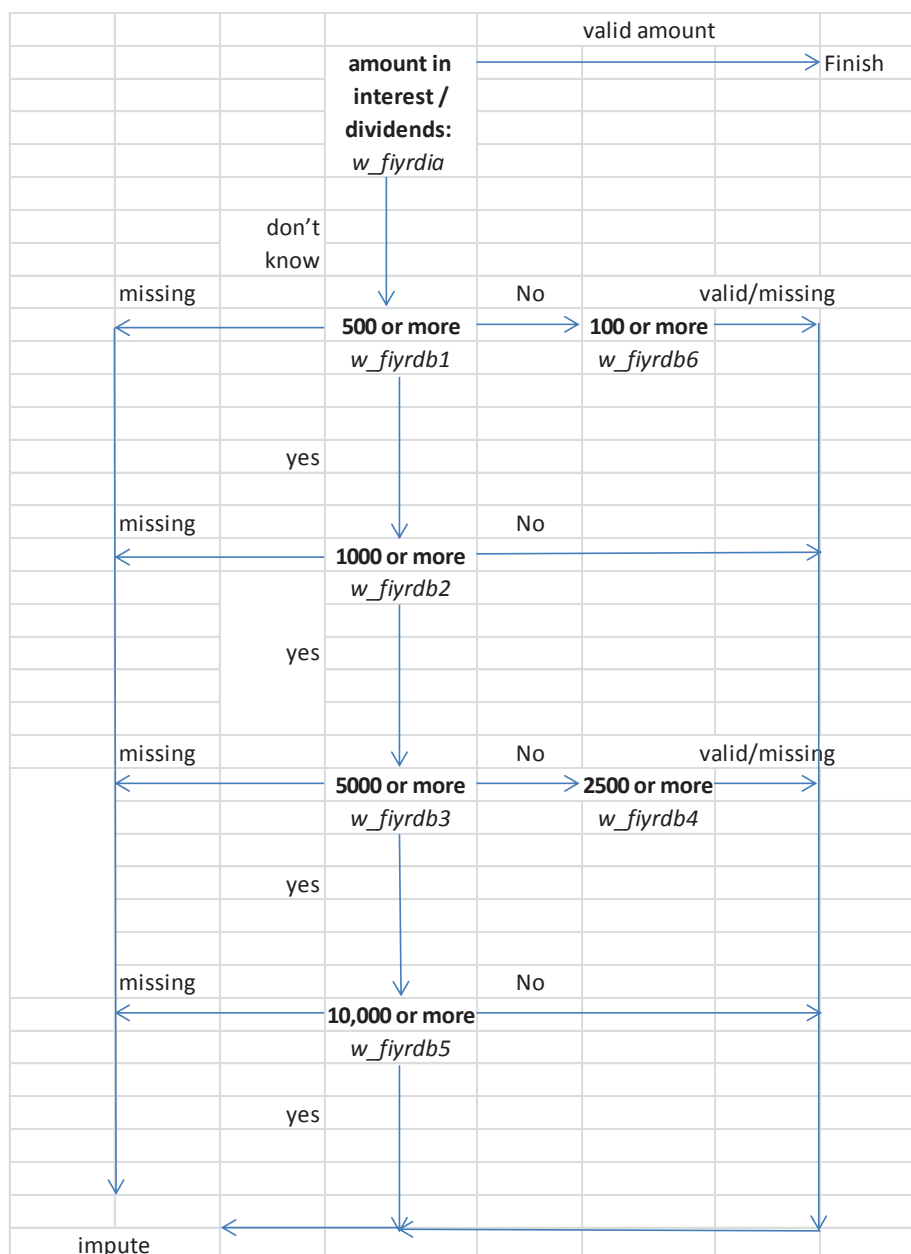
REFERENCES

Avram, S., M. Brewer, P. Fisher, and L. Fumagalli. "Household Income Volatility in the UK, 2009-2017." *Understanding Society Working Paper 2019-07*, 2019.

Benzeval, Michaela, et al. "Understanding Society Innovation Panel Wave 9: : results from methodological experiments." 2017.

Berthoud, R., L. Fumagalli, P. Lynn, and L Platt. *Design of the Understanding Society ethnic minority boost sample.* Colchester: Institute for Social and Economic Research, University of Essex, 2009.

Bollinger, C.R., B.T. Hirsch, C. Hokayem, and J.P. Ziliak. "The Good, The Bad, and The Ugly: Measurement Error, Non-response and Administrative Mismatch in the CPS." *Working paper*, 2018.

Brewer, M., B Etheridge, and C O'Dea. "Why are households that report the lowest incomes so well-off?" *Economic Journal*, 2017: 127, F24–F49.

Brice, J., N., Buck, and E. Prentice-Lane. *(1993). British Household Panel Survey user manual: Volume A: Introduction, technical report and appendices. .* M. F. Taylor (Ed.). University of Essex., 1993.

Budowski, M., R. Tillmann, E. Zimmermann, B. Wernli, A. Scherpenzeel, and A. Gabadinho. "The Swiss Household Panel 1999-2003: Data for research on micro-social change." *ZUMANachrichten* 49, no. 25 (2001).

Burkhauser, R. V., N. Herault, S. P. Jenkins, and R Wilkins. "Survey Under-Coverage of Top Incomes and Estimation of Inequality: What is the Role of the UK's SPI Adjustment?" *Fiscal Studies* 39, no. 2 (2018): 213-240.

Canberra Group. "Handbook on Household Income Statistics, 2nd edition." 2011.

Duncan, G., and M. Hill. "Conceptions of Longitudinal Households: Fertile or Futile?" *Journal of Economic and Social Measurement* 13 (1985): 361-375.

DWP. *Households Below Average Income (HBAI) Quality and Methodology Information Report 2016/17.* Department for Work and Pensions, 2018.

DWP. "Income Dynamics: Income Movements and the Persistence of Low Incomes." Department for Work and Pensions, 2018.

Fisher, P. "Does repeated measurement improve income data quality?" *Oxford Bulletin of Economics and Statistics. doi:10.1111/obes.12296*, 2019.

Francesconi, M., H. Sutherland, and F. Zantomio. "A comparison of earnings measures from longitudinal and cross-sectional surveys: evidence from the UK." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (2011): 297-326.

Frick, J.R., and M. M. Grabka. "Missing Income Data in the German SOEP: Incidence, imputation and its Impact on the Income distribution." 2004.

Frick, J.R., and M.M. Grabka. *Item Non-Response and Imputation of Annual Labor Income in Panel Surveys from a Cross-National Perspective.* IZA DP No. 3043, 2007.

Fumagalli, L., G. Knies, and N. Buck. "Understanding Society, The UK Household Longitudinal Study, Harmonised British Household Panel Survey (BHPS). User Guide." 2017.

Giles, Philip. "An overview of the Survey of Labour and Income Dynamics (SLID)." *Canadian Studies in Population* 28 (2001): 363-375.

Halpern-Manners, A., J. R. Warren, and F. Torche. "Panel conditioning in the general social survey." *Sociological Methods & Research*, 2014: 103–124.

Hayes, C., and N. Watson. "HILDA Imputation Methods." 2009.

Hill, M.S. *The Panel Study of Income Dynamics: a User's Guide. Newbury Park CA: Sage Publications.* Newbury Park CA: Sage Publications, 1992.

Jenkins, S. P. *Changing fortunes: income mobility and poverty dynamics in Britain.* OUP Oxford, 2011.

Kenward, M., and J. Carpenter. "Multiple imputation: current perspectives." *Statistical Methods in Medical Research* 16, no. 3 (2007): 199-218.

Knies, G. *Understanding Society, The UK Household Longitudinal Study Waves 1-7 User Manual.* Colchester: ISER University of Essex, 2017.

Laurie, H. "Continuity and innovation in the design of Understanding Society: the UK Household Longitudinal Study." 2010.

Levy, H., and S.P. Jenkins. *Documentation for Derived Current and Annual Net Household Income Variables, BHPS waves 1-18.* Institute for Social and Economic Research, University of Essex, Colchester, 2012.

Lipps, O. "Income imputation in the Swiss Household Panel 1999-2007." 2010.

Little, and Su. "Item non response in panel surveys." In *Panel Surveys* , by D. Kasprzyk, G. Duncan and Singh. E. panel Surveys. New York: John Wiley, 1989.

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

Lynn, P. *Sample design for Understanding Society.* Understanding Society Working Paper, 2009.

Lynn, P., A. Jäckle, S. P Jenkins, and E Sala. "The impact of questioning method on measurement error in panel survey measures of benefit receipt: evidence from a validation study." *Journal of the Royal Statistical Society Series A* 175 (2012): 289–30.

Lynn, P., A. Nandi, V. Parutis, and L. Platt. "Design and implementation of a high-quality probability sample of immigrants and ethnic minorities: Lessons learnt." *Demographic Research* 38 (2018): 513-548.

Lynn, P., and G. Knies. *Understanding Society Quality Profile. Understanding Society Quality Profile.* Colchester: ISER University of Essex, 2015.

Lynn, P., and G. Knies. *Understanding Society Quality Profile. Understanding Society Quality Profile.* Understanding Society, 2015.

Lynn, P., and M. Borkowska. "Some Indicators of Sample Representativeness and Attrition Bias for BHPS and Understanding Society." Colchester, 2018.

Lynn, P., and O. Kaminska. *Weighting strategy for Understanding Society.* Colchester: ISER, University of Essex, 2010.

Lynn, P., J. Burton, O. Kaminska, G. Knies, and A. Nandi. ""An initial look at non-response and attrition in Understanding Society." Understanding society working paper series 2-2012, 2012.

Lynn, P., S. Jackle, S.P. Jenkins, and E. Sala. "The Effects of Dependent Interviewing on Responses to Questions on Income Sources." *Journal of Official Statistics* 22, no. 3 (2006).

Meyer, B.D., Wallace K, C. Mok, and Sullivan J.X. "Household Surveys in Crisis." *Journal of Economic Perspective* 29, no. 4 (2015): 199-226.

Micklewright, J., and S. V. Schnepf. "How reliable are income data collected with a single question?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2010: 409-429.

Mullan, K., G, Daraganova, and K. Baker. "Imputing income in the Longitudinal Study of Australian Children." 2015.

Postel-Vinayy, F., and A. Sepahsalari. "Labour Mobility and Earnings in the UK, 1992-2016." *Understanding Society Working Paper 2019-09*, 2019.

Ragunathan, E. T., J.M Lepkowski, J. Van Hoewy, and P. Solenberger. "A Multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology* 27, no. 1 (2001): 85-95.

Rubin, D. B. *Multiple imputation for nonresponse in surveys.* New York : Wiley , 1987.

Schafer, J. *Analysis of Incomplete Multivariate Data. .* London: Chapman & Hall, 1997.

Starick, R., and N Watson. "Evaluation of Alternative Income Imputation Methods for the HILDA Survey." 2009.

Starick, R., and N. Watson. "Evaluation of Alternative Income Imputation Methods for the HILDA Survey." 2009.

Van Buuren, S., H.C. Boshuizen, and D.L. Knook. "Multiple imputation of missing blood pressure covariates in survival analysis." *Statistics in Medicine* 18 (1999): 681-694.

Wagner, G., R. Burkhauser, and F. . Bheringer. "The English language public use file of the German Socio-Economic Panel." *Journal of Human Resources 28*, 1993: 429-433.

Webber, M. *Survey of labour and income dynamics: an overview'.* Cat. 75-201E, Ottawa: Statistics Canada: Statistics Canada., 1994.

Westermeier, C., and M.M. Grabka. "Longitudinal Wealth Data and Multiple Imputation. An Evaluation Study." *Survey Research Methods* 10, no. 3 (2016): 237-252.

Wooden, M., S. Freidin, and N. Watson. "The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1." *Australian Economic Review* 35, no. 3 (2002): 339-348.