ENGLISH HOUSING SURVEY (EHS)

Household Dataset
User Guide

2011-12

Contents

Introduction to the English Housing Survey (EHS)
- **Reporting**
- **EHS datasets**
- **Overview of the EHS household dataset**
- **EHS database**
- **Other sources of information**

Using the database - first steps
- **How the household data has been produced**
- **Sample structure**
- **Datasets on the UK Data Archive (UKDA)**
- **Levels of data**
- **Types of data**

Key concepts and definitions
- **Household**
- **Household Reference Person**
- **Un-clustered sampling**

Locating information in the database
- **Interview (household) survey variables**
- **Rotating modules**
- **Multi-coded variables**

How to match files

How to apply grossing factors to obtain national estimates

How to deal with missing data
- **Missing data conventions**
- **Dealing with missing data in analysis**
- **Checking if results are significant**

Appendix A: Content and level of interview survey files

Introduction to the English Housing Survey (EHS)

1 The EHS is a national survey commissioned by the Department for Communities and Local Government (DCLG) that collects information about people's housing circumstances and the condition and energy efficiency of housing in England.

2 The survey has a complex multi-stage methodology consisting of 2 main elements: an initial interview survey of around 13,300 households with a follow up physical inspection of a sub-sample of approximately 6,200 of these dwellings, including vacant dwellings.

3 In 2011-12, a cost review of the EHS resulted in several changes to the survey. The main changes to the survey were:

- a reduction to the EHS sample size
- a reduction in the questionnaire content. This included the removal of the EHS from the Integrated Household Survey (IHS)
- the cessation of the desk based market valuation of sub-sampled properties

Information on these changes can be found in the document "Proposals for changes to the English Housing Survey".

4 The English Housing Survey comprises of:

- an initial interview with approximately 13,300 households a year. This is referred to as the 'full interview sample'. This dataset is made available on a single financial year basis as the EHS household dataset.
- A follow up physical inspection of around 6,200 of these respondents' homes to assess the condition and energy performance of the property. This is referred to as the 'dwelling sample'. This dataset is made available on a rolling two year basis. The 'dwelling sample' comprises all cases where an occupied dwelling has an interview and, physical survey completed plus vacant dwellings with a physical survey only.

5 There is a range of information that will support anyone wishing to make use of the survey data to conduct their own analysis or wishing to understand in detail how the survey is run and managed. All users are encouraged to familiarise themselves with the technical background before undertaking any detailed work using the EHS results.

Reporting

6 There are 3 reports which are published following each survey year. They are:

a) **EHS Headline Report** - the EHS Headline report is usually published around February time. This short report presents preliminary headline findings including key indicators related to departmental housing policies in areas such as trends in tenure and household composition, overcrowding, housing costs, and the condition and energy performance of the stock. Analysis is at a national level from both the full and dwelling samples. This is followed by two separate annual reports, later in the year.

b) **Annual Households Report** - the EHS Households Report is published in the summer. It is based on the full interview sample only and presents comprehensive analysis of housing trends across each of the sectors and for different household groups including changing tenure patterns, overcrowding and under-occupation, rents and mortgages. New topics are introduced on a rolling basis reflecting changing policy interests.

c) **Annual Homes Report** - also published in the summer, the report is based on the rolling two year dwelling sample i.e. cases where a physical inspection has been undertaken. The results presented for '2011' cover the fieldwork period April 2010 to March 2012.

The overarching focus of this report is on housing conditions and energy performance, with other topics introduced each year to reflect emerging policy priorities.

These are available to download from the EHS pages on the DCLG website. The reports include an overview of the survey methodology with guidance about sampling errors and a glossary of terms.

EHS datasets

7 The public datasets that will be available to external users through the UK Data Archive reflect the above EHS reporting strategy. Two separate datasets are therefore available as follows:

a) **EHS Household Dataset** – as used for the EHS Households report. The dataset comprises the full interview data (plus associated derived variables) for all cases where an interview has been completed – 13,300 households per annum (approximately 17,000 per annum before the EHS cost review). Datasets are provided for single financial years together with annual weights. This dataset should be used for any analysis where only information from the household interview is required.

b) **EHS Housing Stock Dataset** – as used for the EHS Homes report. The dataset is available for all cases where a physical survey has been completed and for occupied cases comprises data from the household interview as well as data from the physical survey. For vacant properties only data for the physical survey is provided. The data is made available for a two year rolling sample with the appropriate two year weights. For example, the EHS Housing Stock Dataset is for '2011' covering the period April 2010 to March 2012. This dataset comprises of approximately 14000 cases (previously 16,000 prior to the EHS cost review). The Housing Stock Dataset should be used for any analysis requiring information relating to the physical characteristics and energy efficiency of the housing stock.

8 **This User Guide relates solely to the EHS Household Dataset.** Users who wish to access the EHS Housing Stock Dataset should refer to the separate EHS Housing Stock Data User Guide, which accompanies that dataset. DCLG have deposited the data in this way to provide greater continuity for previous users of the Survey of English Housing (SEH) and English House Condition Survey (EHCS) and so that discrete datasets can be made available in the most straightforward way.

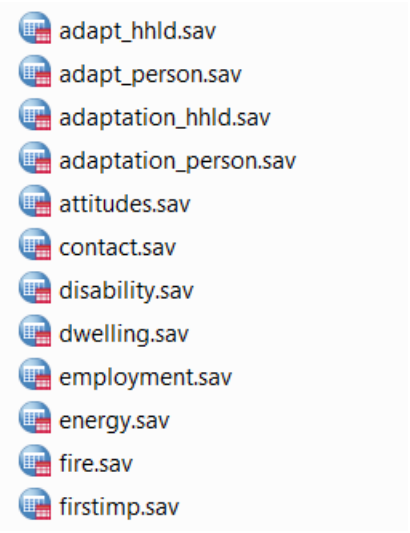
9 Each year new data files are created, some are dropped and a number of new variables are introduced. This applies particularly to the primary data files and users will need to be particularly careful when using these files. Further information is available in the technical background section of the website.

Overview of the EHS household dataset

10 The EHS household data is available in SPSS, STATA and TAB format through the UKDA, as a download. To download or access the data you will need to register with the archive, though it is possible to download descriptions of data and a general overview of the survey without registering.

EHS database

11 Once you have downloaded the EHS data in the preferred format, you will see a number of files within WinZip. The first lot are the **data files**, shown below in SPSS format.

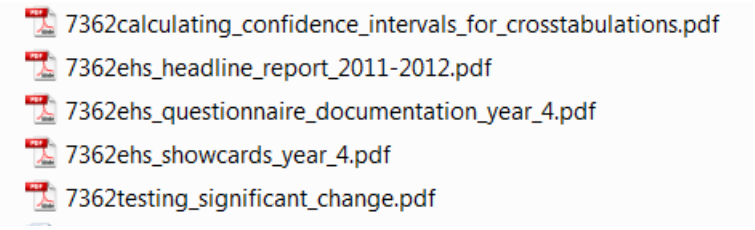


12 The EHS household dataset consists of primary data files containing basic survey data collected by the interviewers. These will be files such as adapt.sav, contact.sav, dwelling.sav and firstimp.sav etc

13 There are also a small number of files containing 'derived' variables on the database. These are holding variables created by calculating with or recoding the basic survey data. These include for example household composition and the bedroom standard, which is used to assess overcrowding and under-occupation.

14 There are two derived files available for the EHS household dataset, which contain the most commonly used derived variables. These are generalfsxx.sav and interviewfsxx.sav where xx is the survey year i.e. 2011-12 will be 11. Users are encouraged wherever possible to make use of these two derived files for their analysis.

15 The second lot of files are the **documentation and reports**

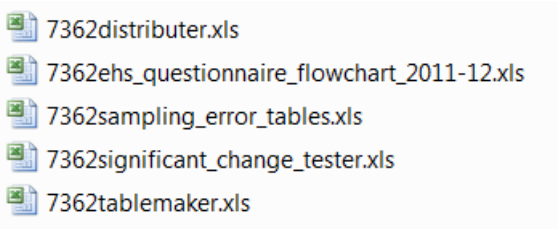


16 A number of documents are included on the EHS household database to help users understand how the database is organised and which data files to use.

17 This includes the **interview survey documentation** which shows all the questions asked of households, the variable names and available response categories and showcards. Some new questions have been introduced and others have been dropped compared to previous years so the 2011-12 interview questionnaire provided indicates the new

interview content. Certain topics are only collected on a rotating basis (see Section 4).

18 The last lot of files on the zip drive are tools to help you with your analysis and afterwards, including how to redistribute missing values and a quick way of checking whether your comparisons are statistically significant.



19 For 2011-12, a new version of the **Data dictionary** has been produced to provide comprehensive details for all the variables on generalfs11.sav, general11.sav, interviewfs11.sav, interview11.sav and physical11.sav files. For information on the variables shown in the primary data files, please refer back to 2010 data.

Other sources of information

20 Further information about the way in which the survey is organised is available on the EHS pages on the DCLG website.

21 As mentioned earlier in this note, all users are encouraged to familiarise themselves with the technical background know as the **EHS Technical Advice Notes** before undertaking any detailed work using the EHS data. These provide details of how the fieldwork is organised, how the sample is drawn and the results grossed each year, estimates of sampling and measurement error together with a full glossary of terms and details of how derived measures such as energy efficiency and income are created.

22 Alongside the three reports and its tables and figures, the EHS also produces **Web Tables**. An updated set of tables are produced each year and made available on the EHS pages on the DCLG website.

Using the database - first steps

23 To use the dataset correctly you must first ensure that you familiarise yourself with and be aware of the following:

- How the dataset has been produced
- What information is published
- The key concepts and definitions used in the survey
- How to locate the information you want in the database (find both survey and computed data by variable and file names)
- How to match files
- How to apply grossing factors to obtain national estimates
- How to deal with missing data
- How to deal with sampling and survey error

Each of these is described briefly in the sections below but please note that this guide is not intended to be a comprehensive source of information. Reference is made to the more detailed explanations and guidance contained in other files, publications and documents which are either at the UKDA or on the DCLG website.

The guide assumes knowledge and experience of using SPSS.

How the household dataset has been produced

24 Primary data are split into a number of topic based files. There are two derived files which contain key variables used in analysis. All files can be matched as required using the case identifier aacode.

Sample structure

Annual sample size	24,299
Frequency of fieldwork	2 months per quarter
Number of months of interview fieldwork	8
Coverage	England only
Clustering	No
Design description	Two-stage sample of addresses
Achieved response rate	62%
Household interviews achieved	13,829
Paired cases achieved	6,459

25 The EHS household sample for 2011-12 was drawn in two stages. This represented a departure from a single stage sample design used in 2010-11. This change was a result of the EHS cost review in 2010-11.

26 An initial sample of 38,416 addresses were drawn from Postcode Address File. These addresses were drawn as a systematic random sample from the Royal Mail's Small User Postal Address File (PAF).

27 For each address, the predominant tenure within the postcode that contained the sampled address was identified and attached to the record. This information was obtained from a commercial dataset[1] held by the Building Research Establishment (BRE) that derives its information from home insurance and mortgage valuation surveys.

Addresses in postcodes that were predominantly owner occupied were sub-sampled with only 54.5% of these addresses being retained in the sample. Social and private rented properties were sampled at a rate of 100%.

As a result, an issued sample of 24,299 addresses was achieved in 2011-12. Further information on this sample design can be found in the English housing survey technical advice note: sampling and weighting – 2011 to 2012 update

28 The allocation of the issued sample continued to follow an un-clustered sample design. An un-clustered sample helps to produce more precise results without increasing the sample size. This is because people with the same characteristics are often geographically clustered. Therefore, by increasing the 'spread' of addresses sampled, this clustering effect is reduced.

29 The principal sampling methodology of the EHS:

- The 2011-12 EHS utilised a two stage sample design. An initial sample of was drawn using a systematic random sample. This was then sub-sampled by predominant tenure, to over-sample private and social rented addresses. This enabled a reduction in the EHS issued sample size, without adversely affecting estimates produced for households in the rented sector. The under-sampling of owner occupiers has been adjusted for in the grossing of EHS estimates.
- The EHS uses an un-clustered sample. This enables a smaller sample to be used with no loss of precision, i.e. without sampling errors being increased. The more scattered sample does, however, have some implications for fieldwork organisation.

[1] Experian possess a database that contains information obtained from a number of sources including insurance companies, Census, etc. referred to as Residata. It is from this that information is taken on predominant tenure within a postcode as well as other information. The matching of the EHS sample to Residata is carried out by BRE.

- The slightly smaller sample achieved in the EHS will give more robust estimates for many measures from the household sample.
- The EHS selects one dwelling per address and one household per dwelling and interviews only the household reference person (HRP) of that household.

30 Full details of sample methodology, and the changes implemented in 2011-12 can be found in the Technical Advice Notes on the EHS pages on the DCLG website

Datasets at the UKDA

31 The EHS data has been accorded National Statistics status. This means the statistics meet user needs, are produced, managed and disseminated to high standards, and managed impartially and objectively in the public interest.

32 The EHS household dataset comprises data collected over a single financial year (April 2011 to March 2012).

The household sample is the set of cases from the interview survey only.

The number of un-weighted cases in the household dataset is as follows:

Household dataset	Un-weighted cases, households	Weighted cases, households (thousands)
EHS 2008-09	**17,691**	**21,530**
EHS 2009-10	**17,042**	**21,554**
EHS 2010-11	**17,556**	**21,893**
EHS 2011-12	**13,829**	**22,040**

Levels of data

33 Most of the data is at household level. However, some of the primary data files contain data at sub-household level. These are indicated at Appendix A. All the files containing derived variables are at household level.

Types of data

34 **Primary survey data** – primary data collected in the field is often very specific and detailed and in many cases is only collected in order to provide the building blocks for computing more useful pieces of information - in particular the household composition and income measures used in the analysis of the survey. There will also be a small amount of missing data in the primary data collected in the field.

In many cases omissions or inconsistencies in the base variables will have been identified in the course of producing derived variables for final analysis. **Users are recommended wherever possible to make use of the final derived variables rather than the raw data from the primary files as these are likely to contain more complete information on key topics such as rent, mortgage payments and income.**

35 **Derived/computed data** – while the information that is collected in the field provides all the base variables needed, a considerable amount of further processing and modelling is needed to fully validate the results and produce secondary variables needed for detailed analysis. Producing derived variables often involves reconciliation of conflicting information from different parts of the survey (e.g. tenure) and/or the imputation of missing data where this has been possible from other data collected in the survey.

36 It is important to note that changes introduced to some of the EHS modules will have a significant impact on any time series analysis attempted and users should familiarise themselves with relevant parts of the Technical Report before drawing conclusions about changes over time. Revised variable names have been used wherever there has been a major break in methodology year on year to minimise the risk of data being misinterpreted.

37 **Missing data** – as with all surveys there will be some level of missing data on the primary data files where information was not available or for example, a respondent refused to answer specific questions. Most of the derived variables for the dwelling sample however have no missing data, as missing values have been imputed based on cases with similar characteristics. Where imputation has taken place there are accompanying flag variables to indicate the level of imputation. Imputation is not always achievable so for some of the key derived variables there will be missing data which you may wish to distribute proportionately using the tool we have provided – see Section 7. The method of data imputation is described more fully in the Technical Advice Notes.

38 Alongside the annual data files, the UK Data Archive holds supporting documentation such as questionnaires, specifications for derived variables and show cards.

39 It should be noted that several variables have been top and/or bottom coded on this dataset. Data has also been cleaned of all identifiable variables to maintain the confidentiality of respondents. Some response categories may also have been condensed for disclosure control reasons.

Key concepts and definitions[2]

40 The household dataset relates to households. It is therefore important to understand how this and related terms are defined and applied in the survey.

41 **Household** – a household is defined as one person or a group of people, who have the accommodation as their only or main residence and (for a group) either share cooking facilities, or share the living accommodation that is a living or sitting room. This represents a slight change in the definition used in the 2010-11 EHS.

42 **Household Reference Person** – the Household Reference Person is defined as the "householder" (that is the person in whose name the accommodation is owned or rented). This term replaced the concept of "head of household" which was used in SEH datasets prior to April 2001. For joint householders (joint owner or joint tenants), the household reference person is whoever has the highest income. If incomes are the same, the older person is defined as the household reference person. Thus the household reference person definition, unlike the old head of household definition, no longer gives automatic priority to male partners.

[2] A full glossary of terms used in the EHS can be found in the EHS Technical Advice Notes and in the Headline and Annual Reports

Locating information in the database

Interview (household) survey variables

43 The topics covered in the Interview Survey questionnaire are illustrated in Figures 1 and 2 below, and include a series of core and household level questions.

44 The core questions used to represent the Integrated Household Survey (IHS) core questions which were asked on the EHS until 2010-11. Although the EHS no longer forms part of the IHS, a series of core questions are still asked on the 2011-12 EHS, due to their relevance to the EHS. These are listed in Figure 1.

45 The EHS collects information on a variety of other topics, some of which were collected on the previous SEH or EHCS. A list of these variables is provided in Figure 2.

Please note that wherever appropriate variable names have remained unchanged from the former SEH or EHCS i.e. when a question has been taken directly from one of the former surveys with the same routing, question wording and response categories the variable name will be unchanged. Where there has been a small change e.g. a change to a response category or a slight change to the routing then a number will have been added to the variable name to indicate this is a variant of the original variable e.g. Vnllrd2 – (From whom did the household rent the property?), previously this was Vnllrd.

Rotating modules

46 Because of pressure on the size of the interview survey, the EHS also uses the concept of 'rotating modules' to bring questions in and out of the survey to meet users' demands and policy needs. The pattern of rotation will vary with some topics being included every other year, while others will be retained for a two year in/two years out basis depending on policy interest. Full details are given on the Interview survey questionnaire.

47 The precise set of files and data provided each year will therefore vary reflecting the changing content of the interview survey. Details of the annual changes are available in the Technical Advice Notes[3].

Multi-coded variables

[3] https://www.gov.uk/government/publications/english-housing-survey-technical-advice

48 The EHS contains a number of multi-coded questions. These are questions where households could give more than one answer; for example 'national identity', 'adaptations for disability' or 'source of funding for buying a property'.

In the EHS household dataset, all multi-coded questions have been stored as a series of binary variables, with a separate variable for each answer option. It is expected that users may find this approach facilitates easier analysis. The following example (Example A) has been taken from the EHS questionnaire documentation:

Example A

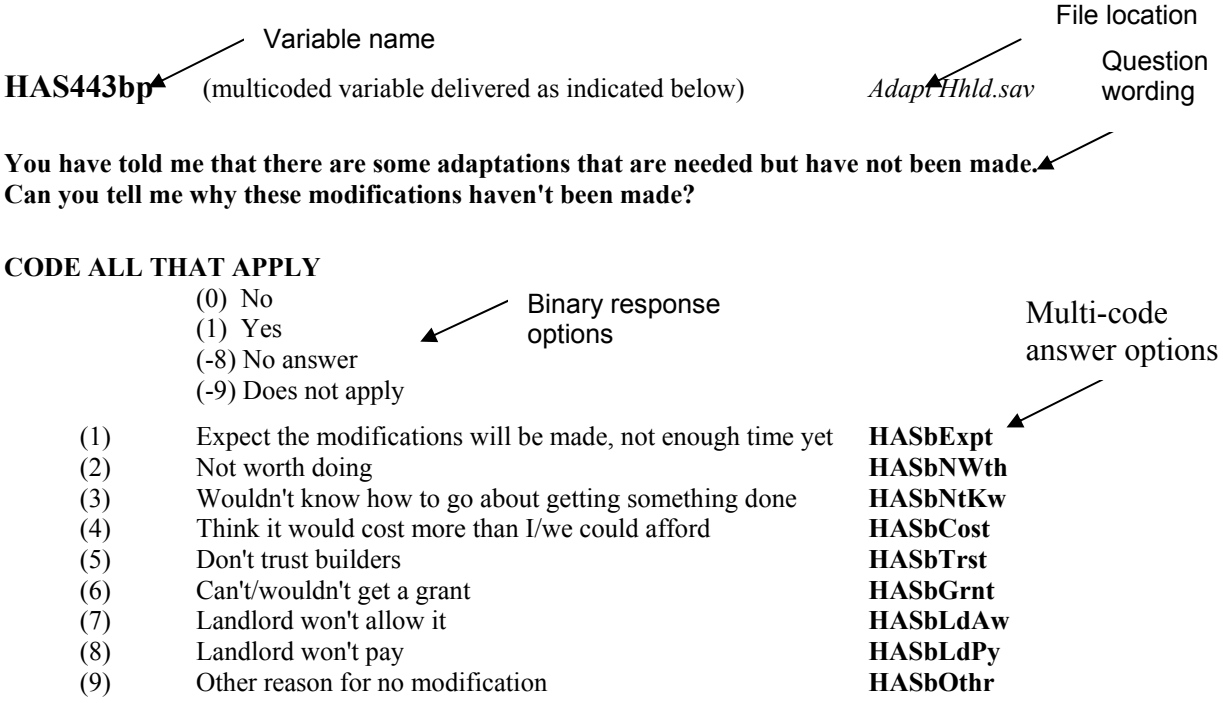
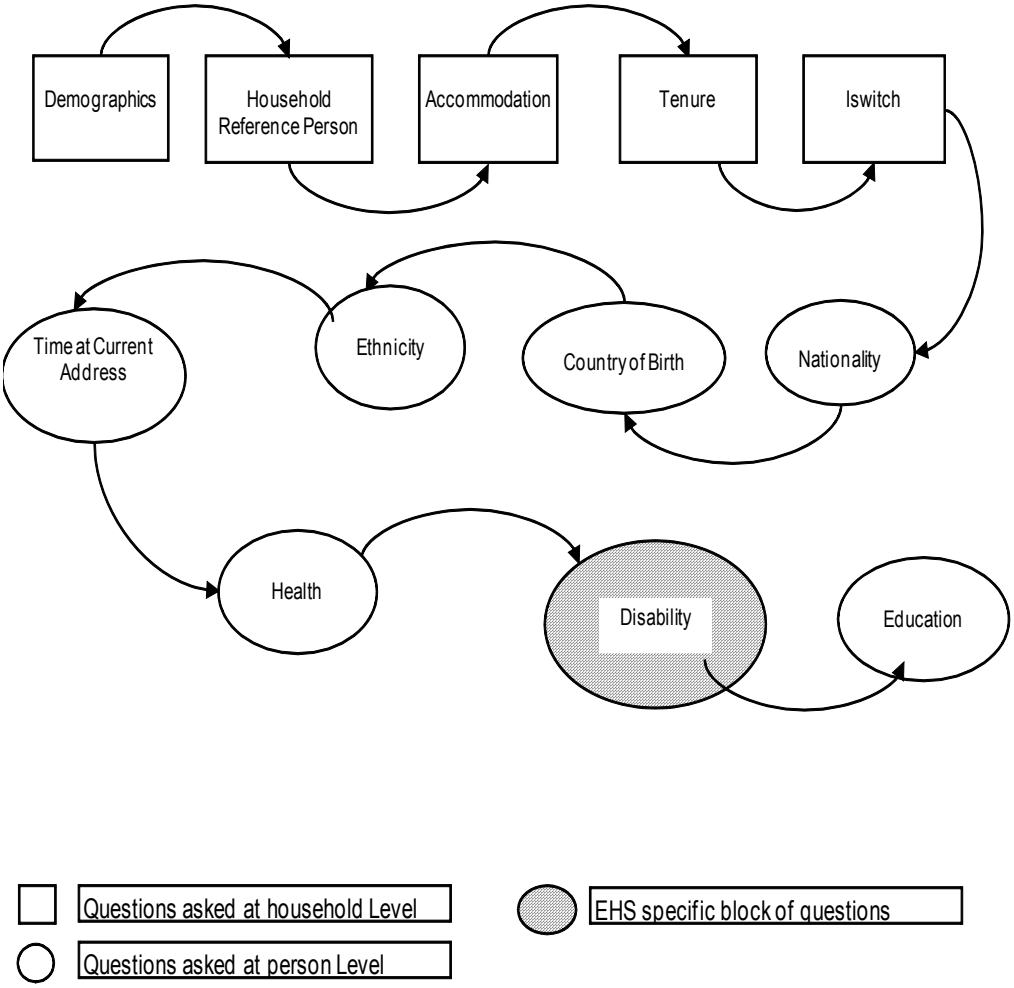
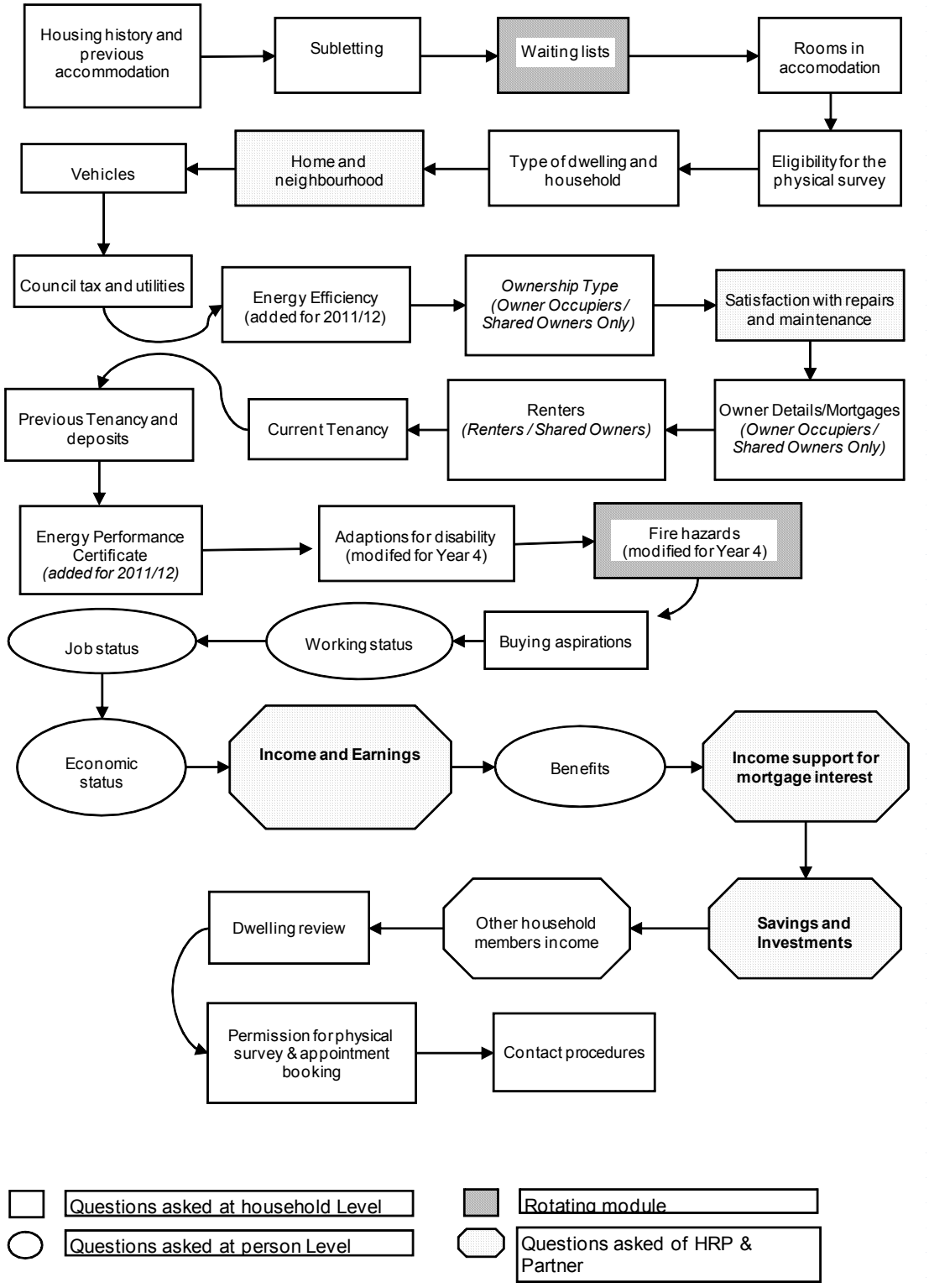


Figure 1: Core questions, EHS 2011-12



Note: The following question topics were dropped in 2011-12: 'National Identity', 'Sexual Identity', 'Religion' and 'Smoking'.

Figure 2: Questionnaire content, EHS 2011-12



Note the following question topics were dropped in 2011-12: 'Dual Nationality', 'Condensation and Damp' and 'Second homes'.

How to match files

Matching files that exist at the household level

49 The key variable for matching at the household level within or across any of the survey data is ***aacode*** (an eight character string variable). You must specify this as the key field when matching. All files on the database have been sorted by this variable so should match. 'Sort file' or 'split file' commands, however, re-sort the file. Matching with standard database files will fail if the data is not first re-sorted by aacode.

Matching lower level files (i.e. below household level) up to the grossing file

50 Files at a lower level of organisation may have more than one case per household. Some of them have exactly the same number of cases per address e.g. the damp.sav file has 6 cases for each address each representing a room for where the question applies. With other files, the number of cases varies e.g. for people.sav the number of cases is the number of people in the household so one person households will just have one case, 6 person households will have 6 cases etc. As there may be more than one case per address, these cannot simply be matched into the household level files to give results at the household level. There are 2 ways to approach these files:

- Match the household file in as ‘tables'. Here the information at household level will be copied for each person etc. within that aacode. The results with then represent the number of people, number of types of window etc. rather than the number of households. This can be useful for very specific analyses, e.g. on numbers of people sharing amenities.
- Aggregate the lower level file up to address code level and then match it in with the household file. (Most key variables for analysis have been derived at the household level so this will not normally be necessary).

How to apply grossing factors to obtain national estimates

51 The EHS household dataset is based on a stratified sample with over-sampling of the rented tenures to achieve large enough sample sizes to produce reliable results for the social and private rented sectors.

Grossing factors have been calculated to:

- Compensate for the design of the sample i.e. the over sampling of some dwellings and under sampling of others; and
- Take account of non-response bias: the survey response rates achieved for different groups of households and dwellings

N.B results must be based on data weighted by the relevant grossing factor to produce national estimates.

52 The EHS household dataset comes with its own grossing factors covering the household sample of cases (see above). The grossing factor can be found in the file **generalfsxx.sav**. The relevant variables are:

- Household grossing – **aagfhxx**

Where xx indicates the year during which fieldwork was conducted. For example, **aagfh11** is the household grossing factor for the 2011-12 household data i.e. for the EHS 2011-12 household dataset.

Aagfhxx should be used for any analysis for which the aim is to provide estimates of households, based on the interview survey data (e.g. percentage of households below the bedroom standard).

The grossing factor can only be used on each full year household datasets. It cannot be used on the dwelling stock data even when this has been split into separate years.

How to deal with missing data

Missing data conventions

53 The files contain no system missing values. Where values are unknown or the question was not asked, specific codes are used and the values are set to user defined missing. The conventions adopted are:

a) Does not apply: -9 for single digit fields, -99 for 2-digit fields etc

b) Unknown: -8 for single digit fields, -88 for 2-digit fields etc

c) Not asked for a particular quarter: -2 for single digit fields, -22 for 2 digit fields etc

d) No partner: -7 for single digit fields, -77 for 2-digit fields etc

54 Users need to be aware of the implications of having missing values switched on or off within SPSS. It is useful for example to leave missing values switched on (the SPSS standard) if users wish to establish the level of missing data in a frequency or cross tab. The SPSS output Case Processing Summary will show the level of missing data. This can be used to decide on whether the level of missing data is significant and whether it is appropriate to re-distribute missing cases.

55 It is particularly important to leave missing values switched on when dealing with a continuous variable e.g. costs to make a home decent. Including -99, -88 etc as a normal value (i.e. not as a missing value) would distort any outputs.

When you create any new variables, you should declare missing values as below:

```
Missing values var1 (-9).
```

This code sets up '-9' as a user defined missing so will be excluded from any cross tabs or statistics of var1.

Dealing with missing data in analysis

56 Analysis that uses derived variables with no missing cases will have no problem with missing data. However, when undertaking analysis using

other variables that have missing or unknown codes, these cases can be redistributed to produce national estimates.

57 Some procedures like 'cross tabulation' exclude missing cases from the table. To include these as the basis for any redistribution it is necessary to switch off the user defined missing values for each variable used. This is done by writing ***'*missing values Var1 ().'** into the SPSS syntax so that missing data will then be included in outputs.

58 Missing values can then be switched back on, by writing: **missing values var1 (-9)** into the SPSS syntax.

59 Where you have missing values within the table, unless there is a clear reason to do otherwise, these need to be allocated pro rata. An Excel spreadsheet **distributer.xls** is included with this database to help to ensure that this is done in a consistent fashion.

Checking if results are significant

60 An additional Excel spreadsheet, **significant change tester.xls**, is included to provide a quick method of checking whether changes in a particular indicator over time are statistically significant. There is an accompanying word document, which explains how the Excel macro works.

Appendix A: Content and level of interview survey files

File name	level	Key identifier	Contents
Adapt Hhld.sav	household	aacode	Information on adaptations made to a home.
Adapt Person.sav	person	persno	Information on adaptations required for someone with a disability.
Adaptation Hhld.sav	household	aacode	Information on the adaptations a household currently has
Adaptation Person.sav	person	persno	Information on adaptations needed for disability purposes
Attitudes.sav	household	aacode	Information on attitudes about current accommodation, area, and other opinion based questions
Contact.sav	household	aacode	Key contact information from the survey
Disability.sav	person	persno	Information on types of disability
Dwelling.sav	household	aacode	Type of accommodation Age of accomodation, housing history Ownership type, buying aspirations
Employment.sav	person	persno	Employment information
Energy.sav	household	aacode	Energy efficiency, Energy Performance Certificate.
Fire.sav	household	aacode	Fires in the accommodation
FirstImp.sav	household	aacode	Key information on the first impressions of the dwelling and neighbourhood as recorded by interviewer
Hhldtype.sav	household	aacode	Tenure, type of dwelling and household, council tax and utilities.
Identity.sav	person	persno	Country of birth, ethnicity and length of residence at current address
Income.sav	household	aacode	Information on amounts and sources of income including benefits
Owner.sav	household	aacode	Ownership details, mortgage
People.sav	person	persno	Key info for each person (age, sex etc.) and relationships with other people in the household
Renter.sav	household	aacode	Renting, social renters, housing benefit, tenancy agreements and deposits,
Rooms.sav	household	aacode	Key information about the number of rooms and whether shared
Vacant.sav	household	aacode	Information relating to vacant properties from the doorstep form
Waitlist.sav	household	aacode	Waiting lists for social housing module