

BACKGROUND AND METHODOLOGY **2010**

CONTENTS

	PAGE
HISTORY OF THE IHS	
SECTION 1 – WHAT IS THE INTEGRATED HOUSEHOLD SURVEY.....	2
 SAMPLE, DESIGN, QUESTIONNAIRE, FIELDWORK AND PROCESSING	
SECTION 2 – SAMPLE DESIGN.....	4
SECTION 3 – THE QUESTIONNAIRE.....	13
SECTION 4 – FIELDWORK.....	14
SECTION 5 – PROCESSING THE DATA - DERIVED VARIABLES AND IMPUTATION.....	16
SECTION 6 – EXPERIMENTAL STATISTICS.....	19
 DATA QUALITY	
SECTION 7 – STATISTICAL QUALITY AND SAMPLING ERRORS.....	21
SECTION 8 – WEIGHTING THE IHS SAMPLE USING POPULATION ESTIMATES.....	25
 PUBLICATION AND DISSEMINATION	
SECTION 9 – IHS DISSEMINATION AND PUBLICATION.....	28

SECTION 1 – WHAT IS THE IHS

The Integrated Household Survey (IHS) is a composite survey combining questions asked on a number of ONS social surveys to produce a dataset of 'Core' variables. The aim of the IHS is to produce high-level estimates for particular themes to a higher precision and lower geographic level than current Office for National Statistics (ONS) Social Surveys.

The Core contains around 100 questions, but a respondent is only asked a proportion of those depending on routing from answers to questions. The Core questions are asked, where possible, at the beginning of the component surveys. There are instances where the Core structure has been modified to accommodate survey specific needs. Every effort has been made to harmonise the Core questions across all IHS surveys, however pragmatic solutions have had to be made in certain cases to ensure surveys still run effectively. The IHS contains questions covering a number of themes. These include:

- Economic Activity
- Education
- Health & Disability
- Identity
- Income

1.1 IHS 2008 – 2009

A set of 'Core' questions were introduced to 3 modules in January 2008; the General Lifestyle Survey (GLF), Living Cost and Food Survey (LCF) and the Opinions Survey (OPN). In April 2008 the IHS Core questions were introduced on the English Housing Survey (EHS) bringing the family of modules on the IHS up to 4.

The data from this period was used as a pilot for the IHS concept, developing the systems and designing the weighting methodology. The IHS data for this period has not been published as it does not provide better quality than existing surveys. The sample size of the IHS for 2008 is significantly smaller than the sample size of the Annual Population Survey, and as the APS carries the majority of the IHS it was deemed the better data source. The only theme where the 2008 IHS data does provide a better quality estimates is for Smoking Prevalence. The APS prior to joining the IHS did not carry this question.

1.2 IHS 2009 – 2010

In April 2009 the IHS Core questions were introduced on the Labour Force Survey (LFS) and Annual Population Survey (APS) questionnaire. Not all the core IHS questions were fully harmonised on the LFS. From June 2009 the Life Opportunities Survey was included in the IHS family of modules.

With the inclusion of these new surveys the IHS became complete, with an achieved annual sample size of approximately 450,000 individuals from interviews undertaken in GB and Northern Ireland. Thereby the first IHS dataset to be released covers the period April 2009 – March 2010, starting the IHS data series from the point that all surveys were included. This large sample size and UK coverage means various geographical breakdowns are possible in the IHS, and it is possible to use a geographical hierarchy to drill down to lower level detail within an area.

The IHS for the period April 2009 – March 2010 contained the following ONS Social Surveys:

- General Lifestyle Survey (GLF),
- Living Cost and Food Survey (LCF)
- the Opinions Survey (OPN),
- English Housing Survey (EHS)
- Labour Force / Annual Population Survey (LFS/APS)
- Life Opportunities Survey (LOS)

In January 2010 the OPN survey was removed from the IHS. This was designed to shorten interview length of the Opinion Survey. The Opinion Survey is entirely financed by sponsors therefore including the survey as a module to the IHS reduced the amount of available space for sponsored questions.

Geographies, on the IHS, include the Countries of the UK, Government Office Region (GOR), Local Authorities (LA) and Nomenclature of Units for Territorial Statistics (NUTS). The latter maintained by Eurostat, the statistical office of the European Communities, as a 3-tier hierarchy used for statistical production across the European Union. The top-level is equivalent to GOR plus England, Scotland, Wales and Northern Ireland, level 2 (NUTS 2) consists of groups of counties, with 37 in the UK, and level 3 (NUTS 3) consists of groups of local authorities with 133 in the UK.

1.3 IHS 2011 – 2012 Component Changes

From April 2011 data from the Life Opportunities Survey (LOS) will no longer be included on the IHS. This is a result of LOS moving to wave 2 interviewing, where the sample is a sub-sample of wave 1 respondents and therefore moving away from a random sample design. These cases are not suitable for inclusion on the IHS.

Also from April 2011 the English Housing Survey will not contain the IHS core questions due to funding restrictions on the EHS.

SECTION 2 – SAMPLE DESIGN

The IHS is compiled from a number of surveys (from this point on they will be referred to as 'modules'). Each of the modules have their own sampling design resulting in the IHS comprising of both clustered and non-clustered, multistage and single stage component samples and also includes cross-sectional and longitudinal data. The sample design of all modules and the detailed implementation are the responsibility of ONS' Social Survey Division.

An overview of the survey design, sample size, methodology, coverage, sample collection and sampling frame for all the modules involved are detailed below within each module sub-heading. They have been grouped into clustered and un-clustered survey designs. More details and in-depth background on individual surveys can be found here:

APS - www.statistics.gov.uk/StatBase/Product.asp?vlnk=1537&Pos=&ColRank=1&Rank=272

LOS - www.statistics.gov.uk/about/methodology_by_theme/life-opp-survey.asp

EHS - www.communities.gov.uk/publications/housing/ehcstechrep2005

GLF - www.statistics.gov.uk/StatBase/Product.asp?vlnk=5756

LCF - nswebcopy/statbase/Product.asp?vlnk=361

OPN - www.ons.gov.uk/about/who-we-are/our-services/omnibus-survey/index.html

2.1 CLUSTERED/UNCLUSTERED SAMPLING

Originally, the intention was to make the sampling designs for all module unclustered sampling. However, due to budget constraints, time and different systems, it was decided that the sampling designs for the modules will remain as they were pre-IHS mode.

Unclustered Sample Definition: The design means that addresses are sampled directly from the entire Postcode Address File (PAF) rather than being selected from a sub-group of postcode sectors. As a result, addresses in any quota may be more dispersed than those on another survey which uses a clustered sample design. The dispersion of addresses will depend upon the population density in each particular region. For example, a city centre is likely to be more densely populated; therefore the quota is likely to cover a smaller geographic area than a rural area which is likely to have a lower population density. An unclustered 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.

Clustered Sample Definition: Due to a smaller number of field interviewers, clustered design sampling is used as it enables coverage of the population without having all the addresses available from the Postcode Address File. This is because similar characteristics from certain areas can be grouped together. From the remaining postal sectors available, n postal sectors are randomly selected and within them, m number of addresses are selected to go into the sample.

This method of sampling and the consequent weighting affect the sampling errors for the module estimates. The effect can be shown by calculating the Effective Sample Size which gives the size of an equal probability sample, equivalent in precision to the unequal probability sample actually used.

2.2 ONE AND MULTI-STAGE SAMPLING

Based on whether the module is clustered or unclustered determines whether a one-stage sampling or multi-stage *stratified* sampling is being used.

As mentioned earlier, unclustered sample design means that addresses are sampled directly from the entire Postcode Address File (PAF)

For clustered sampling, the PAF file is initially grouped into Government Regions – these would be strata and postal sectors are selected from systematically in each stratum – this would be the first stage of the sampling. Within each postal sector, postcodes are then randomly selected, giving the second-stage sampling.

2.3 UNCLUSTERED MODULES

2.3.1 APS¹

Sample Size	334,206
Proportion of IHS sample (%)	74
Frequency	Monthly
No. of months of data in the IHS dataset	12
Coverage	UK
Clustering	No
Design Description	Single stage sample of addresses (delivery points)
PSU	Address
SSU	N/A

The Annual Population Survey (APS) is a combined survey of households in United Kingdom. Its purpose is to provide information on key social and socio-economic variables between the ten-yearly censuses, with particular emphasis on providing information relating to small geographical areas.

The APS comprises the Quarterly Labour Force Survey (LFS) (waves one and five for the UK), plus data from the Annual Local (Area) Labour Force Survey (LLFS) Boosts for England, Scotland and Wales.

¹ More information on the APS can be found at [www..statistics.gov.uk/StatBase/Product.asp?vlnk=1537&Pos=&ColRank=1&Rank=272](http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=1537&Pos=&ColRank=1&Rank=272)

APS dataset structure

Time		Q2 2009	Q3 2009	Q4 2009	Q1 2010
LFS cases	cohort 1	wave 5			
	cohort 2	wave 4	wave 5		
	cohort 3	wave 3	wave 4	wave 5	
	cohort 4	wave 2	wave 3	wave 4	wave 5
	cohort 5	wave 1	wave 2	wave 3	wave 4
	cohort 6		wave 1	wave 2	wave 3
	cohort 7			wave 1	wave 2
	cohort 8				wave 1
	cohort 9				
	cohort 10				
	cohort 11				
	cohort 12				
APS (boost) cases	cohort a1	wave 4			
	cohort a2	wave 3			
	cohort a3	wave 2			
	cohort a4	wave 1			
	cohort a5				
	cohort a6				

The APS including the APS Boost aimed to ensure that a sample of at least 510 economically active persons in English Local Authority Districts (except London Boroughs where the target is 450 economically active persons) was achieved.

The size of the total APS achieved sample is 334,206 people with roughly 53% of the sample from waves 1 and 5 of the LFS and 47% from the APS (boost).

The main dissemination route for labour market data from the APS is Nomis[®]. Other data from the APS, for example, on housing, education, employment, ethnicity and health is available from the Neighbourhood Statistics and National Statistics websites. Datasets are also available at the UK Data Archive and from the LFS Customer Service (SOS) team in ONS.

2.3.2 LOS²

Sample Size	23,368
Proportion of IHS sample (%)	5
Frequency	1 month per quarter
No. of months of data in an IHS dataset	4
Coverage	GB
Clustering	No
Design Description	Single stage sample of addresses (delivery points)
PSU	Address
SSU	N/A

LOS started with a baseline random sample of 37,500 households across Great Britain, interviewing all people aged 16+ in each household. This baseline survey started in June 2009 and will take two years to complete. At March 2010, first half of baseline interviews were completed.

The sample size of 37,500 households is necessary to achieve an adequate sample of the British population and key sub-groups within it. In particular, a sample of this size will allow life opportunities to be differentiated by type of impairment and other diversity characteristics. For the baseline survey, all people aged 16+ in the household will take part in face to face interviews. Parents or guardians will also be asked to provide some key data about children, aged 11 to 15. Children will be able to take part in a face to face interview when they reach age 16.

The longitudinal design enables three distinct groups to be followed over time:

1. Disabled group
2. Comparison group of non-disabled people
3. A larger non-disabled group, monitored for the onset of impairment over time.

Respondents will be interviewed every 12 months, whether they are resident in the original household or have since moved. Following the baseline survey only the first two groups will be interviewed face to face. The third group will be monitored over time via a shorter telephone interview to check for the onset of impairment. If any member of the household has experienced the onset of impairment, they will be invited to take part in a face to face interview.

The LOS Wave 1 interviews will be conducted across Great Britain over two years. All people aged 16 and over who are resident at the sampled address will be interviewed in person. Questions about children aged 11-15 will be answered by parents. All respondents to Wave 1 will be followed up over time using both face to face and telephone interviewing.

² More information on the LOS can be found at www.statistics.gov.uk/about/methodology_by_theme/life-opp-survey.asp

2.3.3 EHS³

Sample Size	40,753
Proportion of the IHS sample (%)	9
Frequency	2 months per quarter
No. of months of data in the IHS dataset	8
Coverage	England
Clustering	No
Design Description	Single stage sample of addresses (delivery points)
PSU	Address
SSU	N/A

Originally, 32,100 addresses were selected as a systematic random sample from the PAF ordered by postcode. This systematic sampling ensured that the sample's distribution across Local Authorities (LAs) is close to the overall PAF distribution.

This unclustered sample was one of the key challenges in the survey design (which has clear benefits in terms of statistical and cost efficiency) and the requirement for the surveyors to work in fairly contained areas for the physical surveys. Due to the delays in implementing the unclustered sample design planned for ONS' Integrated Household Survey (IHS), an alternative field work design was developed for the first year of the EHS increasing the sample size to 40,753.

It was decided to make use of existing LFS fieldwork geography and overlay a new geography for the EHS. Allocation of the sampled addresses to fieldwork quarters was carried out to achieve 'clustering' of addresses within LFS interviewer areas.

³ More information on the EHS can be found at www.communities.gov.uk/publications/housing/ehcstechrep2005

2.4 CLUSTERED MODULES

2.4.1 GLF⁴

Sample Size	18,033
Proportion of IHS sample (%)	4
Frequency	Monthly
No. of months of data in an IHS dataset	12
Coverage	GB
Clustering	Yes
Design Description	Multi-stage stratified random sampling
PSU	Postcode Sector
SSU	Address

The GLF sample design follows a four-year sample rotation in which people remain in the sample for four years (waves) with one quarter of the sample being replaced each year. Each quarter of the sample is known as a replication, and each replication is representative of the target population. Each year one of the four replications is dropped and a new one added, giving an overlap of 75 per cent between successive years. This has been implemented to avoid high attrition rates due to repeated interviewing.

At the end of the interview, GLF ask the wave 1, wave 2 and wave 3 respondents for their agreement to be re-interviewed in the following year. Agree to recall rate for wave 1, 2 and 3 adults is around 84%.

Sample Replication	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
1	1st						
2	1st	2nd					
3	1st	2nd	3rd				
4	1st	2nd	3rd	4th			
5		1st	2nd	3rd	4th		
6			1st	2nd	3rd	4th	
7				1st	2nd	3rd	etc.
8					1st	2nd	etc.
9						1st	etc.

	New sample
	Follow-up sample

For the initial interview (Wave 1) all addresses are selected at random from the 'small user' Postcode Address File covering England, Scotland and Wales.

For follow-up quotas (Waves 2, 3 and 4) GLF want to re-interview all adult household members aged 16 and over to see if their situation has changed since they were last interviewed. New adult household members are also interviewed i.e. anyone who has moved in with an Original Sampled

⁴ More information on the GLF can be found at www.statistics.gov.uk/StatBase/Product.asp?vlnk=5756

Member (OSM), and the 15 year olds who have now become 16 years old and therefore eligible for a personal interview. Any household member who refused last time the rest of the household were interviewed, or was a non-contact, is still eligible to be interviewed this time.

Since GLF wants to follow the individuals from the original sample, movers will need to be traced and, if still in scope, interviewed in their new home. Any adults in their new household are also eligible for interview.

2.4.2 LCF⁵

Sample Size	11,989
Proportion of IHS sample (%)	3
Frequency	Monthly
No. of months of data in the IHS dataset	12
Coverage	GB
Clustering	Yes
Design Description	Multi-stage stratified random sampling
PSU	Postcode Sector
SSU	Address

The LCF sample for Great Britain is a multi-stage stratified random sample with clustering. It is drawn from the Small Users file of the Postcode Address File - the Post Office's list of addresses. All Scottish offshore islands and the Isles of Scilly are excluded from the sample because of excessive interview travel costs. Postal sectors are the primary sample unit.

The provisional overall response rate for the 2009 LCF was 52.5 per cent in Great Britain.

5,019 households in Great Britain co-operated fully in the survey in 2009; that is they answered the household questionnaire and all adults in the household answered the full income questionnaire and kept the expenditure diary.

A further 222 households provided sufficient information to be included as valid responses.

Each individual aged 16 or over in the household is asked to keep a detailed record of expenditure every day for two weeks. Children aged between 7 and 15 are also asked to keep a simplified diary of daily expenditure. Information provided by all members of the household is kept strictly confidential. Each person aged 16 and over in the household who keeps a diary (and whose income information is collected) is subsequently paid £10 high street voucher, children who keep a diary are given a £5 voucher.

The fieldwork is conducted by the Office for National Statistics (ONS) in Great Britain, and by the Northern Ireland Statistics and Research Agency (NISRA) of the Department of Finance and Personnel in Northern Ireland, using almost identical questionnaires. Households at the selected addresses are visited and asked to co-operate in the survey.

⁵ More information on the LCF can be found at nswebcopy/statbase/Product.asp?vlnk=361

2.4.3 OPN⁶

Sample Size	20,981
Proportion of IHS sample (%)	5
Frequency	Monthly
No. of months of data in the IHS dataset	9
Coverage	GB
Clustering	Yes
Design Description	Multi-stage stratified random sampling
PSU	Postcode Sector
SSU	Address

The Opinions survey uses what is known as random probability sampling broken into Government Office Region, proportion of households with no car, socio-economic classification and the proportion of people aged over 65 years.

In common with most other ONS surveys, the Opinions survey draws its sample from the Royal Mail's Postcode Address File (PAF) of 'small users'. This PAF contains the addresses for approximately 27 million private households in the UK which receive fewer than 50 items of mail per day. It is the most up-to-date and complete address database in the UK. Each month 67 postal sectors are selected, with probability of selection proportionate to size. Within each sector, 30 addresses are chosen randomly giving a final sample of 2,010 addresses each month. At the start of the interview, the interviewer determines the household composition and then he or she selects the respondent from among all those aged 16 and over. This selection is performed at random using a Kish grid.

All household members over the age of 15 years are asked a set of classificatory questions but only one person per household is selected to answer the Opinions module questions. Proxy responses are not permitted on the Opinions modules. As only one person per household is interviewed, the data are subsequently weighted to correct for the unequal probability of selection that this causes. Applying these weights will gross up the data by age, sex and region to the population control totals used on the Labour Force Survey (LFS). As well as accounting for the unequal probability of selection, these weights correct for certain types of non-response bias and improve precision for most variables. Weights will be supplied at person level in each survey month and, if required, will be available at a household level and on a quarterly or annual basis.

This method of sampling and the consequent weighting affect the sampling errors of the survey estimates. The effect can be shown by calculating the Effective Sample Size which gives the size of an equal probability sample, equivalent in precision to the unequal probability sample actually used. The Effective Sample Size will vary slightly from one month to another with the proportions of interviews in different sized households. On average the Effective Sample Size of the Opinions survey is 84 per cent to 86 per cent of the actual sample of individuals. An achieved sample of 1800 individual adults in the Opinions survey is equivalent to an equal probability sample of about 1500.

⁶ More information on the OPN can be found at <http://www.ons.gov.uk/about/who-we-are/our-services/omnibus-survey/index.html>

When the sample for a particular month is selected by the ONS's Sampling Implementation Unit, a letter is sent out to the sampled address approximately one week before the start of the field period. This letter explains the purpose of the survey and advises the residents that an interviewer will be calling regarding the survey. As an incentive to respond to the survey, a book of postage stamps is included with the letter. The interviewer is required to call at all sampled addresses to ensure that they are residential and not business addresses. Businesses, institutions, temporary accommodation and vacant addresses are ineligible in most instances. The Opinions Survey interviews are conducted using Blaise questionnaire programming tool on a laptop and, in most instances, the interviewer enters the respondents' replies directly into the computer. If the questions are considered to be sensitive then respondents are given an option of self completion.

From beginning to end the survey period lasts 14 weeks. This includes four weeks for developing and testing the questionnaire in Blaise (a computer-assisted survey design and processing system). The survey will be in the field for 4-5 weeks, including time for telephone reissues (the Office for National Statistics (ONS) Telephone Unit re-contact initial non-contacts and refusals). During the final five weeks of the survey period the data will be cleaned, weighted, analysed and tabulated by the research team.

SECTION 3 – THE QUESTIONNAIRE

3.1 MANAGEMENT OF THE IHS QUESTIONNAIRE

The IHS questionnaire content is determined by ONS. ONS are responsible for identifying, in conjunction with other government departments, needs for new questions or changes to existing questions (e.g. changes in legislation or new government employment programmes) and for determining priorities, given the constraint of interview length. Where possible, ONS, has ensured that harmonised questions and harmonised concepts are included in the IHS. ONS also ensures that European Union data requirements are met.

There is the opportunity for Government Departments to sponsor questions on the IHS. Currently Communities and Local Government and the Department of Health sponsor the IHS questions.

3.2 GOVERNANCE OF THE IHS QUESTIONNAIRE

It is essential that changes to the IHS core are rolled out on the IHS modules in a systematic way to ensure the Core remains harmonised and reduces the potential for 'gaps' in the IHS dataset being formed. The IHS team in ONS oversees this process on a day to day basis.

The initial point of contact for all requests for changes will be with the IHS team. Depending on the type of change requested the IHS Team will decide the appropriate action required. The types of change are categorised as:

- Minor wording change to question or interviewer instructions
- New categories added to questions
- Significant changes to wording or routing change- resulting in change of context to the question.
- New core questions

Minor changes and such as altering wording and addition of new categories are approved by the IHS team if they do not change the meaning of the question. Any changes to harmonised questions are discussed at the National Harmonisation Working Group.

Changes that significantly alter the meaning of a question or the addition of new questions are considered at the ONS Social Survey Design Authority and also at the IHS Steering Group (which has representatives from Government Departments and other interested bodies). Changes are implemented following agreement at these bodies.

All changes, except for issues which have significant impact on survey outputs (i.e. missing questions etc) are implemented in January. New questions are tested following the LFS cognitive testing programme, so new questions will have to be provided to the IHS Team a year in advance of the field period they are to be asked in.

SECTION 4 – FIELDWORK

4.1 THE CONDUCT OF FIELDWORK

4.1.1 FACE-TO FACE AND TELEPHONE INTERVIEWING

IHS fieldwork is carried out by the interview field force which is comprised of both face-to-face and telephone interviewers. The face-to-face interviewers work from their home while telephone interviewers, work in a centralised Telephone Unit in Titchfield, Hampshire. At the start of 2009, there were approximately 1000 interviewers working on the IHS; 800 in the field and 200 in the telephone unit.

4.1.2 FIELDWORK DOCUMENTS

Advance letters and purpose leaflets that are given to selected household do not contain any information regarding the IHS. The fieldwork documents are module specific.

4.2 REQUIREMENTS FOR ANSWERS TO QUESTIONS

Whilst every effort is made to obtain answers to all relevant questions from each respondent, it is recognised that there will be some cases when a respondent genuinely does not know the answer to a particular question (particularly in the case of responses by proxy - see below) and cases when a respondent does not wish to give the answer to a particular question. In general ONS would not wish to lose such respondents and a "no answer" or "don't know" will be accepted.

However, there are a number of key questions in the survey, some of which are fundamental in classifying a respondents' economic status, which, if not answered cause that whole record (though not the whole household) to be dropped.

4.2.1 FORCED RESPONSE QUESTIONS

These 'forced response' questions are currently as follows:

R1-16	Relationship to head of household and to other household members
SEX	Sex of respondent
DVAGE	Age of respondent
MARSTA ⁷	Marital status
MARCHK	Is spouse a member of household
LIVWTH	Whether respondent is living together with someone as a couple
HRPID	Whether accommodation is owned/rented in respondent's name
SCHM08	Whether respondent on a government scheme in the reference week
NDTYPE4	New Deal participants
NEWDEA4	For those respondents on New Deal, the type of option in which they were participating
TECLEC	Whether respondent was on a TEC, LEC or some other scheme
WRKING	Whether respondent did any paid work in the reference week
JBAWAY	Whether respondent was away from a paid job in the reference week
OWNBUS	Whether respondent did any unpaid work in the reference week for a business owned by him/herself
RELBUS	Whether respondent did any unpaid work for a business owned by a relative
EVERWK	Ever had a paid job or place on scheme
STAT	Whether respondent was working as an employee or self-employed

⁷ In the Blaise questionnaire this question is XMARSTA.

LOOK4	Whether respondent was looking for paid work in the previous 4 weeks
LKYT4	Whether respondent was looking for a place on a Government scheme in the previous 4 weeks

4.3 PROXY INTERVIEWS

4.3.1 ACCEPTABILITY OF PROXY RESPONSES

All surveys allows interviewers to take answers to questions by proxy if a respondent is unavailable. This is usually from another related adult who is a member of the same household, although there are exceptions to this rule:

- (i) a young person, of the same household, may translate for a non-English speaking relative;
- (ii) a carer, of the elderly or infirm, although not related, may answer for someone in their care if it can be established that they know the respondent well enough;
- (iii) anyone can respond by proxy with the personal permission of the head of household or spouse.

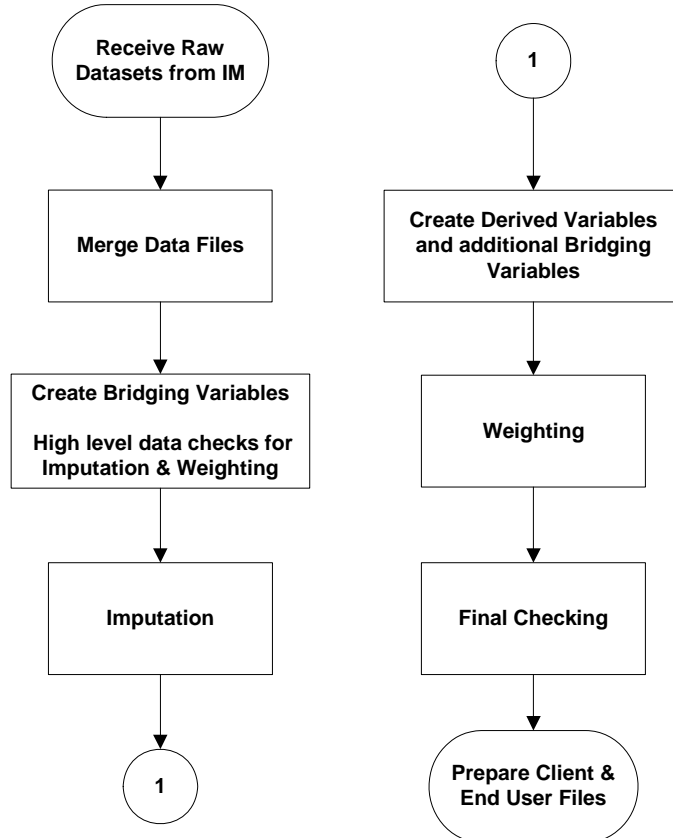
4.4 COMPUTER ASSISTED PERSONAL INTERVIEWING (CAPI)

The IHS interviews are carried out by face-to-face interviewers using laptop computers and by telephone interviewers using networked desktop microcomputers. The questionnaire and edit instrument that they use is identical in both modes. It is produced using the BLAISE CAPI software package, which was created by Statistics Netherlands. SSD has designed a computer system which takes the output from BLAISE and uses it to create derived variables, to weight up population estimates and other processes leading to the production of data files for customers. The system also rotates the data for use at the next wave of interviewing.

SECTION 5 – PROCESSING THE DATA – DERIVED VARIABLES AND IMPUTATION

5.1 UNWEIGHTED RAW DATASETS

Unlike other surveys, the IHS dataset is not received at the aggregated quarterly level. This means the module datasets have to be merged before being able to process the data. This is all done within the Research Team in ONS Social Survey. Processing the data include cleaning and editing the data relevant for the IHS.



5.2 DERIVED VARIABLES

In order to analyse IHS data, a number of derived variables (DVs) are specified. DVs are variables which combine the answers to two or more questions from the questionnaire. All DVs are specified for the UK as they are created after the merge of all individual survey data. Volume 3 contains the latest set of flow diagrams used to specify the current DVs, which can be split into four groups:

- Person:** DVs created for each individual record on the database;
- Family:** DVs created for each family on the database; on a flat file each family member would carry the same value for each family based DV;
- Household:** DVs created for each household on the database; on a flat file each household member would carry the same value for each household based DV.

5.3 BRIDGING VARIABLES

Similar to derived variables, bridging variables are also created. These new variables are created when variable names and options vary from one module to another. Examples of bridging variables include:

- The IHS question SOLO is asked if the respondent is or has been self-employed. APS goes one step further and splits the results into two variables: SOLOR for current job and SOLOLR for last job. For the IHS dataset, SOLOR and SOLOLR has been merged and the results are added to SOLO to create SOLODV.
- The IHS STAT question is asked for employment status. The APS variables STATR and STATLR cover this for current job and last job respectively. These have been merged and added to STAT to create STATDV.
- The IHS DWELLTYP is a derived variable from ACCOM, HSETYPE, FLTTYP and ACCOTH. However, the APS has a non-derived variable which includes this question. The APS DWELLTYP has been merged with IHS' DWELLTYP.

5.4 HARMONISED DATA

Bridging is currently used as the core questions have not all been harmonised. These will be amended at a later date.

5.5 DATA CHECKING

Whilst some checking is performed in-the-field by the BLAISE survey instrument, other checks are carried out once the data have been received back from interviewers in the field or from the telephone unit.

The principles of the checks are to ensure that the data have no duplication of records etc., that the data have the correct household structure, in terms of persons in the household, and that certain key variables have valid values. These checks are important in maintaining the quality of the data. Examples include:

- checking that families have been correctly assigned within households by referring to the relationship grid
- country checks to ensure geographies are consistent across variables
- checks to ensure each household has a household reference person (HRP)
- checks to ensure that there is a record for every member of a household – known as the 'full house check'
- checks to ensure that the household type (hhtype) matches the household composition as defined in the relationship grid
- checks on ages for outliers (e.g. very old, old workers etc)

5.6 IMPUTATION

Like many other surveys, the IHS core dataset can contain multiple levels of non-response. These can be at:

- (i) household level - non-response for a complete household;
- (ii) individual level - non-response for one or more individuals in a responding household;

- (iii) item level - incomplete response for an individual.

In this section, the main concern is with the second level where at least one individual has not responded in an otherwise responding household. If a complete household level interview is made, then all core data for non-responding individuals in the household can be imputed for. As a by product of the process, any item level missing values will also be imputed for.

The planned imputation process broadly follows the approach developed for the 2011 Census. That is, to impute the person level variables in a series of modules, specified by topic, at the household level. The strength of this approach is that it follows a joint imputation method, within module, which utilises information from the other household members and also preserves the multivariate distributions of the data. Thus, the imputation process is being developed in a series of eight modules which align to the core topics as follows:

1. Nationality, Country of Birth, Ethnicity
2. Period at current address
3. Health and Smoking
4. Income
5. Government Training Scheme
6. Employment
7. Looking for work
8. Education

In order to find a donor household that is as similar as possible to the recipient, variables that are related to the missing values are identified to act as matching variables. We seek to identify strong relationships between the target variables to be imputed and those used to select the donors. There are a number of standard methods which include: logistic regression, building regression trees (for example SPSS AnswerTree), imputing within categorisation of the output variables. The first two methods are appropriate when there are a small number of target variables, for example in the imputation of the APS where the key target variable related to ILO status (ILODEFR). However, where there are a large number of target variables, it is important to identify a small subset of predictors to act as matching variables through out the process. This also ensures that the imputation remains robust to changes overtime amongst the multivariate distributions.

In terms of IHS, the complete information from the household level interview is being used to specify the matching variables. To date these include: Gender, Working Age, ILO Status, NS-Sec and Relationship to HRP.

SECTION 6 – EXPERIMENTAL STATISTICS

6.1 WHAT ARE EXPERIMENTAL STATISTICS?

These are statistics that are in the testing phase and are not fully developed.

6.2 WHEN ARE STATISTICS EXPERIMENTAL?

Defining what is experimental and non-experimental is largely a matter of statistical judgement, but typically experimental series arise when:

- they are being produced part way through a well defined development programme - whether these statistics are new or changed versions of existing statistics;
- statistics are new but still subject to testing in terms of their volatility and ability to meet customer needs;
- the statistics do not yet meet the rigorous quality standards of National Statistics; or
- a rich variety of new measures is available from a new set of statistics, with components that have considerable immediate value to users. These users are aware of the statistics' theoretical quality and can make use before ONS has completed all operational testing. The testing is designed to fully validate the measures to the standard expected of National Statistics.

6.3 WHY PUBLISH IHS EXPERIMENTAL STATISTICS?

The aims of the publication include:

- Consultation: ONS would like to get informed feedback from potential users
- Acclimatisation: where the experimental series are alternative versions of the format of existing series, ONS may wish to help users become accustomed to new presentations.
- Use: experimental series can provide very useful information for users as long as their nature is well explained and understood.

6.4 HOW SHOULD USERS INTERPRET SUCH SERIES?

It is important that users clearly understand the status of these series and the cautions that apply (which will vary for each statistic).

6.5 WHEN DO EXPERIMENTAL DATA BECOME NON-EXPERIMENTAL?

This will be a matter of statistical judgement, but typically will take into account factors such as:

- When it is judged that statistical methods have settled down
- When coverage reaches a good level
- When user feedback indicates that these statistics are useful and credible
- When the defined development phase has ended
- When it is judged that the statistics meet the rigorous quality standards of National Statistics.

6.6 POLICY ON DISSEMINATION OF EXPERIMENTAL STATISTICS

The 2009-2010 series will be introduced by an article on the NS Website and in the appropriate "Trends" publication, to explain their concepts and construction etc.

Experimental series will then be updated on the NS Website. They will be accessible via a dedicated area of the Latest Releases section. This will ensure that the status of the series is clear. Articles and/or tables can be made available for downloading.

Whilst updates for experimental series will be disseminated via the Website, the information may also appear in paper publications, such as the next edition of the appropriate "Trends" publication. It is important that in all cases these statistics are clearly marked as experimental.

SECTION 7 – STATISTICAL QUALITY AND SAMPLING ERRORS

7.1 DIMENSIONS OF QUALITY

There are six dimensions of quality, when it comes to ensuring how useful the data is going to be. They are:

- Relevance
- Accuracy
- Timeliness and Punctuality
- Accessibility and Clarity
- Comparability
- Coherence

The table below gives an overview of what the key components are. More details can be found on in the ONS Guidelines for Measuring Statistical Quality⁸.

Definition	Key Components
Relevance	
The degree to which the statistical product meets user needs for both coverage and content.	Any assessment of relevance needs to consider: <ul style="list-style-type: none"> • who are the users of the statistics; • what are their needs; and • how well does the output meet these needs?
Accuracy	
The closeness between an estimated result and the (unknown) true value.	Accuracy can be split into sampling error and non-sampling error, where non-sampling error includes: <ul style="list-style-type: none"> • coverage error; • non-response error; • measurement error; • processing error; and • model assumption error.
Timeliness and Punctuality	
Timeliness refers to the lapse of time between publication and the period to which the data refer. Punctuality refers to the time lag between the actual and planned dates of publication.	An assessment of timeliness and punctuality should consider the following: <ul style="list-style-type: none"> • production time; • frequency of release; and • punctuality of release.
Accessibility and Clarity	
Accessibility is the ease with which users are able to access the data. It also relates to the format(s) in which the data are available and the availability of supporting information. Clarity refers to the quality and sufficiency of the metadata, illustrations and accompanying advice.	Specific areas where accessibility and clarity may be addressed include: <ul style="list-style-type: none"> • needs of analysts; • assistance to locate information; • clarity; and • dissemination.

⁸ www.statistics.gov.uk/downloads/theme_other/Guidelines_Subject.pdf

Comparability	
The degree to which data can be compared over time and domain.	Comparability should be addressed in terms of comparability over: <ul style="list-style-type: none"> • time; • spatial domains (e.g. sub-national, national, international); and • domain or sub-population (e.g. industrial sector, household type).
Coherence	
The degree to which data that are derived from different sources or methods, but which refer to the same phenomenon, are similar.	Coherence should be addressed in terms of coherence between: <ul style="list-style-type: none"> • data produced at different frequencies; • other statistics in the same socio-economic domain; and • sources and outputs.

7.2 ACCURACY

The following section focuses on the ‘accuracy’ dimension which is composed of sampling errors and non-sampling errors.

- non-sampling errors may be important, but are often difficult to measure. ONS attempt to minimise non-sampling error through a number of techniques including interviewer training, weighting adjustments to minimise the impact of non-response, to name a few.
- sampling errors arise because we observe only a sample and not the whole population. The theory of sample design allows us to estimate the size of the sampling errors for the IHS. This appendix provides measurements of sampling errors for the IHS.

7.3 SAMPLING ERRORS

An important component of statistical quality is sampling error, which arises because the variable estimates are based on a sample rather than a full census of the population. The results obtained for any single sample would be likely to vary slightly from the true values for the population. The difference between the estimates derived from the sample and the true population values is referred to as the sampling error. The amount of variation can generally be reduced by increasing the size of the sample, and by improving the sample design. Standard errors have been measured for estimates derived from the Integrated Household Survey (IHS), and these may be used to assess the accuracy of the estimates presented in this report.

7.3.1 CALCULATING STANDARD ERRORS

Unlike non-sampling errors, it is possible to estimate the size of the sampling error. The standard error (se) of a percentage p , based on a simple random sample of size n is calculated by the formula,

$$se(p)_{srs} = \sqrt{\frac{p(100-p)}{n}}$$

The April 2009 to March 2010 IHS comprises three multi-stage surveys (LCF, OPN, GLF) and three single-stage surveys (APS, EHS, LOS). The complexity of these designs means that

sampling errors calculated on the basis of a simple random sample design will not reflect the true variance in the survey estimates.

Although the largest component of the IHS comes from unclustered surveys (APS, EHS and LOS) the other surveys (LCF, OPN and GLF) have a clustered design. For the purpose of selecting the sample for each of the clustered surveys, the respective sampling frame is stratified by GOR-based area identifiers (which exist on the sampling frame) together with certain Census Area Statistics which are matched to the sampling frame. For the LCF and GLF, there is explicit stratification by 26 GOR-based regions, NSSEC and car ownership. The OPN sample is explicitly stratified by 30 GOR-based regions, NSSEC and number of pensioners. For the APS, EHS and LOS there is no explicit stratification in the sampling design, but the fieldwork allocation strategies for these surveys cause an implicit stratification by interviewer area. When calculating the standard errors, it is therefore necessary to take both clustering and stratification into consideration.

Clustering can lead to a substantial increase in sampling error if the households or individuals within the primary sampling units (PSUs) are relatively homogenous but the PSUs differ from one another. By contrast, stratification tends to reduce sampling error and is particularly effective when the stratification factor is related to the characteristics of interest on the survey.

All the IHS component surveys use systematic sampling to select units (delivery points). However, the multistage surveys use postcode sector groups as primary sampling units (PSUs), whereas the single stage surveys use delivery points (addresses) as the PSU. Standard error calculations for systematic surveys are typically more complex than for other random selection procedures. Our standard approach in ONS is to pair off contiguous PSUs into collapsed strata and to base the variance of the estimator on the squared differences between PSUs within strata, summed over strata. It would not be appropriate to mix the postcode sector PSUs from multistage samples with those of single stage samples of households, therefore PSUs in the IHS are paired so that PSUs within a given collapsed stratum always belong to the same survey.

Because of the complexity of the IHS sample design, the size of the standard error depends on how the characteristic of interest is spread within and between the PSUs and strata. The method used to calculate the standard errors for the survey takes this into account. It explicitly allows for the fact that the estimated values (percentages and means) are ratios of two survey estimates: the number with the characteristic of interest is the numerator (y) and the sample size is the denominator (x), both of which are subject to random error.

The standard error of a survey estimate is found by calculating the positive square root of the estimated variance of the ratio. The formula used to estimate the variance of a ratio r (where $r = y/x$) is shown below.

$$\text{var}(r) = \frac{1}{x^2} [\text{var}(y) + r^2 \text{var}(x) - 2r \text{cov}(y, x)]$$

$\text{Var}(r)$ is the estimate of the variance of the ratio, r , expressed in terms of $\text{var}(y)$ and $\text{var}(x)$ which are the estimated variances of y and x , and $\text{cov}(y, x)$ which is their estimated covariance. The resulting estimate is only valid if the denominator (x) does not vary too greatly. The method compares the differences between totals for adjacent PSUs (postal sectors in the case of LCF, OPN, GLF; addresses in the case of APS, EHS, LOS) in the characteristic of interest. The ordering of PSUs reflects their ranking on the explicit or implicit stratifiers used in the sample design.

7.3.2 DESIGN FACTORS

The design factor, or *deft*, of an estimate p is the ratio of the complex standard error of p to the standard error of p that would have resulted had the survey design been a simple random sample of the same size.

$$deft(p) = \frac{se(p)}{se_{srs}(p)}$$

This is often used to give a broad indication of the effect of the clustering on the reliability of estimates. The size of the design factor varies between survey variables reflecting the degree to which a characteristic of interest is clustered within PSUs, or is distributed between strata. For a single variable the size of the design factor also varies according to the size of the subgroup on which the estimate is based, and on the distribution of that subgroup between PSUs and strata. Design factors below 1.0 show that the complex sample design improved on the estimate that we would have expected from a simple random sample, probably due to the benefits of stratification. Design factors greater than 1.0 show less reliable estimates than might be gained from a simple random sample, due to the effects of clustering. Design factors equal to 1.0 indicate no impact of the survey design on the reliability of the estimate relative to a simple random sample.

The formula to calculate the standard error of the difference between two percentages for a complex sample design is:

$$se(p_1 - p_2) = \sqrt{\left[deft_1^2 \left(\frac{p_1(100 - p_1)}{n_1} \right) + deft_2^2 \left(\frac{p_2(100 - p_2)}{n_2} \right) \right]}$$

where p_1 and p_2 are observed percentages for the two sub-samples and n_1 and n_2 are the sub-sample sizes.

7.3.3 CONFIDENCE INTERVALS

The estimate produced from a sample survey will rarely be identical to the population value, but statistical theory allows us to measure its accuracy. A confidence interval can be calculated around the estimated value, which gives a range in which the true value for the population is likely to fall. The standard error measures the precision with which the estimates from the sample approximate to the true population values and is used to construct the confidence interval for each survey estimate.

The 95% confidence intervals have been calculated for each estimated value presented. These are known as such, because if it were possible to repeat the survey under the same conditions a number of times, we would expect 95% of the confidence intervals calculated in this way to contain the true population value for that estimate. When assessing the results of a single survey, it is usual to assume that there is only a 5% chance that the true population value falls outside the 95% confidence interval calculated for each survey estimate. To construct the bounds of the confidence interval, 1.96 times the standard error is subtracted from, and added to, the estimated value, since under a normal distribution, 95% of values lie within 1.96 standard errors of the mean value. The confidence interval is then given by:

$$p \pm 1.96se(p)$$

The 95% confidence interval for the difference between two percentages is given by:

$$(p_1 - p_2) \pm 1.96se(p_1 - p_2)$$

If this confidence interval includes zero then the observed difference is considered to be a result of chance variation in the sample. If the interval does not include zero then it is unlikely (less than 5% probability) that the observed difference could have occurred by chance.

SECTION 8 – WEIGHTING THE IHS SAMPLE USING POPULATION ESTIMATES

8.1 WEIGHTING THE IHS

8.1.1 SUMMARY

Population weighting serves two purposes. First it enables tables showing population estimates to be produced, second it compensates for differential non-response among different sub groups in the population.

The IHS collects information on a sample of the population. To convert this information to give estimates for the population we must weigh the data. Each case is given a weight which can be thought of as the number of people that case represents. In a perfect world each person in the population would have an equal probability of being selected for the IHS (as in a simple random sample). The weight would simply be the reciprocal of this probability. However, because of differential non-response, some people are more likely to be in the sample than others. People with a lower probability of being in the sample, such as young people in London should have a higher weight.

It is impossible to measure directly what probability each member of the sample had of being selected. Instead, the population is split into sub-groups (or cells) where the number of people in each sub-group is known (based on population estimates). The known population weights are calculated by assigning each case in the sub-group the weight calculated by dividing the population in that subgroup by the number of cases in the sample in that subgroup. As a result the weighted estimates of the total population for that sub-group equals the actual population and the weighted estimate of the total population from the sample will equal the known estimate.

There is one limitation to this method. If the cells are too small, there is a possibility that none of the sample will come from that subgroup. The weighted estimate of the population in that subgroup will be zero and hence the estimate of the total population will be too small. To avoid this (and to make the task of producing the population control totals easier) a multi-stage population weighting procedure is used. Each stage corrects for a different cause of non-response: stage 1 corrects for non response at a local area level, stage 2 corrects for non-response amongst young people by age and sex; and stage 3 corrects for non-response by region, age (banded) and sex.

8.2 DETAIL OF THE IHS WEIGHTING PROCEDURE

8.2.1 COMPONENTS OF THE IHS WEIGHTS

The weight for each case (person) is a product of the following:

- The initial address-level design weight
- An adjustment weight for multi-household addresses
- A non-response adjustment
- An attrition adjustment (where applicable)
- A scale factor for pooling surveys
- A calibration adjustment.

8.2.2 THE INITIAL ADDRESS-LEVEL DESIGN WEIGHT

For both the multi-stage surveys and the single-stage surveys, address selection probabilities are calculated using the same mathematical formulae as those currently used in the respective stand-

alone surveys. The design weight for each address is calculated as the reciprocal of the address selection probability. The use of the reciprocal of the selection probability as a design weight is non-controversial, as it is the standard Horvitz-Thompson estimator.

8.2.3 ADJUSTMENTS FOR MULTI-HOUSEHOLD ADDRESSES

Where an address contains more than one household an adjustment weight for multi-household addresses is computed. This is the reciprocal of the conditional probability of selecting the chosen household from all households at the multi-occupied address. Large multi-household adjustments are winsorized. Winsorization replaces extreme values by an upper-bound that is less extreme. This helps to reduce the variance of the weights.

8.2.4 NON-RESPONSE ADJUSTMENTS

Non-response adjustments are applied to the cross-sectional surveys and the first waves of the longitudinal surveys. Attrition adjustments are applied for waves greater than one. The proposal to use non-response weighting to adjust for potential self-selection bias introduced by the response mechanism is based on sound logic and is often employed where appropriate auxiliary data are available for both respondents and non-respondents. To be effective, non-response adjustments must adjust for the situation where response rates differ between classes and non-responders tend to have different characteristics from responders. Otherwise we run the risk of simply adding variance to the weights, which increases the variance of the estimator, without any gain in bias reduction, i.e. the mean square error (MSE) is increased.

The LCF and GLF modules both use non-response weighting based on the decennial census non-response link, which results in fixed (assumed internally homogeneous) weighting class adjustments, which are assumed to remain constant during the intercensal period. LFS does not use this adjustment because no self-selection bias was found during the last census non-response link study. For reasons of user clarity and system cohesion we use the same method of non-response adjustment for all surveys in the IHS (i.e. replacing the census non response link classes for GLF and LCF).

We use the ONS Output Area Classification (OAC), which has 53 groups describing the socio-demographic make-up of small geographic areas, as our auxiliary data source. It is our belief that OAC is likely to be related to a number of the IHS core variables and, as such, will be useful in reducing bias in the estimator. The design weighted response/non-response binary indicator is regressed onto the classes of the OAC and GOR. We can then distinguish whether response propensity varies by OAC class and use the reciprocal of the response rates in the relevant OAC classes as non-response weighting adjustments. Non-response adjustment weights were calculated for LCF, OPN, EHS, LOS, GLF wave 1, APS wave1 and APS newly found households. The initial non-response adjustment was calculated as the reciprocal of the predicted response probability.

8.2.5 ATTRITION ADJUSTMENT

A further non-response adjustment is required for longitudinal surveys. Attrition is the term given to distinguish non-response after the first interview from non-response to the first interview. As with initial non-response, attrition has the potential to introduce bias through differential rates of drop-out. However, as there is usually a wealth of detail on such cases the loss to follow up in successive waves from data collected in the previous wave(s), the potential to correct for potential bias is much greater. It is common procedure to adjust longitudinal surveys for attrition. We acknowledge that only limited attrition adjustment data is available on the IHS core. However, data for constructing suitable attrition models is obtained by using data from previous waves of the standalone surveys. Therefore, for the APS attrition model, characteristics of APS households were found from APS datasets for previous waves. Likewise, for the GLF attrition model, characteristics of GLF persons and households were found from GLF datasets for previous

waves. Attrition adjustment weights were calculated as the reciprocal of the predicted probability that the case responds in the current wave given that it was a responding case in the previous wave.

The attrition adjustment follows the same logic as the adjustment for initial non-response, i.e. calculating response propensities from a logistic model and using the reciprocal of these as an attrition adjustment. It is these attrition adjusted weights (i.e. the multiplicand of the attrition weight and the previous wave's non-response/attrition-adjusted design weight) which then get entered as the initial weight for calibration at time t.

8.2.6 POOLING THE SURVEYS

To pool the six surveys into a single survey (the IHS), the adjusted design weights were scaled using the three steps outlined below.

Step 1: pooling the un-clustered surveys

First, the un-clustered surveys are scaled relative to their achieved sample sizes within local authority. Their design weights are adjusted so that, within each local authority, the combined total of the scaled design weights d'_{ijk} summed to the local authority population.

Let

- i be an index representing an un-clustered survey which is part of the IHS (i is one of APS, EHS or LOS).
- j be an index representing a local authority.
- k represent the k^{th} individual in survey i and local authority k .
- d_{ijk} be the adjusted design weight for the k^{th} individual belonging to survey i and local authority j . All individuals in the same household have the same value for d_{ijk} because d_{ijk} is the address-level design weight multiplied by the product of the multi-household adjustment weight, the non-response adjustment weight and the attrition adjustment weight for the household to which that individual belongs.
- n_{ij} be the achieved sample size for survey i in local authority j .
- N_j be the population total for local authority j , which is supplied from ONS Centre of Demography (ONSCD).

Then, the formula used to calculate the design weights for pooling the un-clustered surveys is given by

$$d'_{ijk} = \left(\frac{n_{ij}}{\sum_{i \in \{\text{un-clustered surveys in IHS}\}} n_{ij}} \right) \frac{d_{ijk} N_j}{\sum_{\ell=1}^{n_{ij}} d_{ij\ell}}$$

Step 2: scaling the clustered surveys

Each clustered survey is treated separately. For each clustered survey, the design weights are adjusted so that, within each government office region (GOR), the combined total of the scaled design weights d'_{ijk} summed to the GOR population.

Let

- i be an index representing an clustered survey which is part of the IHS (i is one of GLF, LCF or OPN)
- r be an index representing a GOR (each local authority j is completely nested within a GOR r).
- M_r be the population total for GOR r , which is supplied from ONSCD.

Then, for each clustered survey, each design weight is scaled by the GOR population total divided by the sum of all design weights within that GOR and survey.

$$d'_{ijk} = \frac{d_{ijk} M_r}{\sum_{\{j:j \in r\}} \sum_{\ell=1}^{n_{ij}} d_{ij\ell}}$$

Step 3: combining the clustered and un-clustered surveys

First, the un-clustered surveys are calibrated as a single survey using Statistics Canada Generalised Estimation System (GES) software and supplying d'_{ijk} as the a-weight. The calibration constraints used for the un-clustered surveys are chosen to be the same as those used for APS integrated household calibration weighting. At the end of the calibration process, the variance of the resulting calibration weights, within GOR, are calculated and saved.

Next, each clustered survey is calibrated separately using GES software and supplying d'_{ijk} as the a-weight. The calibration constraints used for the clustered surveys were chosen to be similar to those used for each respective stand-alone survey. At the end of each calibration process, the variance of the resulting calibration weights, within GOR and survey, are calculated and saved.

Finally, the IHS dataset is partitioned into weighting classes determined by survey and GOR (government office region), where the un-clustered surveys are treated as a single survey. Denote the variance of the calibration weights within GOR and survey by v_{ir} where i is one of GLF, LCF, OPN, and 'UNC' where 'UNC' represents the un-clustered surveys pooled together. The final dataset is constructed by adjusting the weights within each weighting class in inverse proportion to the standard error of its calibration weights. The scale factor is

$$a_{ir} = \frac{\left(\frac{1}{\sqrt{v_{ir}}}\right)}{\sum_i \left(\frac{1}{\sqrt{v_{ir}}}\right)}$$

The final scaled design weight is $d^*_{ijk} = a_{ir} d'_{ijk}$ for individual k in local authority j within GOR r and the i^{th} survey. This is the design weight which is used for all further calibration steps in the IHS weighting process.

8.2.7 THE INTEGRATED HOUSEHOLD WEIGHT

Integrated household weights are then produced by calibrating the scaled, adjusted household design weights (d_{ijk}^*) to gross the entire IHS dataset to the UK population. The IHS contains details of all household members containing data on their age, sex and region, which are the variables used in the calibration model. It also imputes data for all missing adults. Consequently, when we use integrative calibration, which results in all individuals in the household having the same weight as the household weight, we reproduce the population totals by summing the weights within the calibration control classes. As there are no missing individuals in the dataset, estimators are unbiased (assuming the non-response adjustments have worked correctly).

Calibration has a number of desirable properties.

- It introduces consistency within known population estimates.
- It potentially reduces bias associated with variables included in the calibration model.
- It often reduces the variance of estimates.

To produce the integrated household weights, calibration weighting was carried within a Generalised Regression (GREG) framework using the GES. The control groups used for integrated household calibration on the IHS are as follows.

Calibration group 1: Sex **by** age.

Sex (the gender of the respondent)

1 = Male 2 = Female

Age (five year age band with those 75 years and older as a single group)

1 = "0-4"	5 = "20-24"	9 = "40-44"	13 = "60-64"
2 = "5-9"	6 = "25-29"	10 = "45-49"	14 = "65-69"
3 = "10-15"	7 = "30-34"	11 = "50-54"	15 = "70-74"
4 = "16-19"	8 = "35-39"	12 = "55-59"	16 = "75+"

Calibration Group 2: Local Authority code (2009 definition)

In order to speed up the computation, the data was partitioned by government office region. Cases from each government office region in England and cases for Wales, Scotland and Northern Ireland were calibrated separately.

8.2.8 THE SEXUAL IDENTITY WEIGHT

For the sexual identity question, a weight is needed which is applied only to cases responding to the sexual identity question i.e. not responses collected by proxy. For non-proxy adults, the scaled, adjusted household design weights are calibrated to gross this subset of the IHS dataset the UK adult population. For the sexual identity question which is specific to adults and for which proxy response is not accepted, this calibration step makes a non-response adjustment for the proxies within the selected calibration groups. The adjustment puts a larger weight on the non-proxy responses received from persons in those calibration groups.

In the Apr09-Mar10 dataset, the sexual identity question was asked of all adult respondents in the survey. However, the sexual identity question was not asked of proxy respondents. If the proxy indicator for the interview in Apr09-Mar10 was set to 'true', then this indicated that the corresponding case was to be omitted from the sexual identity weighting and from the sexual identity analysis.

The persons included in the sexual identity weighting for Apr09-Mar10 dataset were those persons who fit the criteria ((age>=16) and (proxy=1)), i.e. those 16 years and above who did not have a proxy interview. Note –proxy variable was renamed sidproxy to harmonise to the second quarter dataset onwards (see below).

From the second quarter of 2010, the routing in the APS questionnaire was changed so that the sexual identity question was no longer asked of APS adults who had answered this question in a previous period. The idea was to reduce respondent burden by using the fact that it is unlikely that a persons sexual identity will change between one wave of the APS and the next, thereby making it possible to roll forward the sexual identity responses from the previous period. Under this new situation, the proxy identifier associated with the current APS interview only corresponded to the current proxy identifier if the APS respondent was being asked the sexual identity question in the current period. It was necessary then, to create a separate proxy identifier for the sexual identity question. This variable was called sidproxy and constructed so that sidproxy is equal to the current proxy identifier if the sexual identity response occurred in the current quarter, but sidproxy is equal to the previous proxy identifier if the sexual identity response was brought forward from the previous quarter.

The persons included in the sexual identity weighting for Jul09-Jun10 dataset and later periods will be those persons who fit the criteria ((age>=16) and (sidproxy=1)), i.e. those 16 years and above who did not have a proxy interview.

Calibration for the sexual identity weight was carried out using GREG methodology implemented within the GES. In order to speed up the computation, the data was partitioned by country and government office region. Cases from each government office region in England and cases for Wales, Scotland and Northern Ireland were calibrated separately. The population totals used for calculating the non-proxy adult weight and the sexual identity weight are population estimates for persons 16+ within each calibration group. Calibration groups used are for each region is shown below. They are very similar to the controls used on APS person weighting, with the exception that here (for the purpose of calculating the non-proxy adult weight and the sexual identity weight) they are only being applied to persons aged 16 years and above.

Calibration controls used for Government Office Regions in England

Calibration Group 1: England & Wales grossing area **by Sex by age** (for persons 16+ in the selected government office region in England).

England & Wales grossing area (England was divided into 45 grossing areas. These grossing areas are groupings of local authorities which are similar to each other. The grossing areas used here are those nested within the selected government office region)

Sex (the gender of the respondent)
 1 = Male 2 = Female

age (these are five year age band for persons in England 16+, with those 75 years and older as a single group).

4 = "16-19"	5 = "20-24"	9 = "40-44"	13 = "60-64"
	6 = "25-29"	10 = "45-49"	14 = "65-69"
	7 = "30-34"	11 = "50-54"	15 = "70-74"
	8 = "35-39"	12 = "55-59"	16 = "75+

Calibration Group 2: England Local Authorities aggregated by respondents of working age **by English combined Local Authorities** (for persons 16+ in the selected government office region in England).

England Local Authorities aggregated by respondents of working age 1 = 'Not working age (under 16 years)'

2 = 'Female - working age (16-59)'

3 = 'Male - Working age (16-64)'

English combined Local Authorities (these are individual Local Authorities in England, with the exception that the two local authorities London and Westminster are combined as a single group).

Calibration Group 3: GOR **by** Age (0 – 23 years) **by** Sex (for persons 16+ in the selected government office region in England).

GOR (the selected government office region)

Age (single year of age for persons 16-24 and persons 25+ grouped together).

1 = "0-15"	7 = "21"
2 = "16"	8 = "22"
3 = "17"	9 = "23"
4 = "18"	10 = "24"
5 = "19"	11 = "25+"
6 = "20"	

Sex (the gender of the respondent)

Calibration controls used for Wales

Calibration Group 1: England Wales grossing area **by** Sex **by** age (for persons 16+ in Wales).

EW grossing area (Wales was divided into 10 grossing areas. These grossing areas are groupings of Welsh local authorities which are similar to each other.)

Sex (the gender of the respondent)

1 = Male 2 = Female

Age (these are five year age band for persons in Wales 16+, with those 75 years and older as a single group).

4 = "16-19"	5 = "20-24"	9 = "40-44"	13 = "60-64"
	6 = "25-29"	10 = "45-49"	14 = "65-69"
	7 = "30-34"	11 = "50-54"	15 = "70-74"
	8 = "35-39"	12 = "55-59"	16 = "75+"

Calibration Group 2: Local Authority **by** working age grouped with sex

Local Authority (these are local authorities in Wales)

Work age grouped with sex (an age by sex grouping)

1 = "Male 0-15"	9 = "Female 0-15"
2 = "Male 16-24"	10 = "Female 16-24"
3 = "Male 25-34"	11 = "Female 25-34"
4 = "Male 35-44"	12 = "Female 35-44"
5 = "Male 45-54"	13 = "Female 45-59"
6 = "Male 55-64"	14 = "Female 60-74"
7 = "Male 65-74"	15 = "Female 75+";
8 = "Male 75+"	

Calibration Group 3: Age (0-23 years) **by** Wales NUTS⁹ group by Sex

Age (single year of age for persons 16-24 and persons 25+ grouped together).

1 = "0-15"	7 = "21"
2 = "16"	8 = "22"
3 = "17"	9 = "23"
4 = "18"	10 = "24"
5 = "19"	11 = "25+"
6 = "20"	

Wales NUTS group (these are Nomenclature of Units for Territorial Statistics level 2 codes for Wales – NUTS2)

Sex (the gender of the respondent)

Calibration controls used for Scotland

Calibration Group 1: Scotland Glasgow Local Authority **by** Sex **by** Age (25 groups)

Scotland Glasgow Local Authority (indicator for whether or not the local authority is within Glasgow)

Sex (the gender of the respondent)

Age (a 25-category age grouping)

1 = "0-4"	11 = "23"	21 = "65-69"
2 = "5-9"	12 = "24"	22 = "70-74"
3 = "10-15"	13 = "25-29"	23 = "75-79"
4 = "16"	14 = "30-34"	24 = "80-84"
5 = "17"	15 = "35-39"	25 = "85+"
6 = "18"	16 = "40-44"	
7 = "19"	17 = "45-49"	
8 = "20"	18 = "50-54"	
9 = "21"	19 = "55-59"	
10 = "22"	20 = "60-64"	

Calibration Group 2: Local Authority **by** Sex and Age Scottish Local Authority aggregate

Local Authority (these are local authorities in Scotland)

Sex and Age Scottish Local Authority aggregate (Small local authorities in Scotland Clackmannanshire (00QF), Orkney (00RA), Shetland (00RD) and Eilean Siar Western Isles (00RJ) are labelled "smLA" and are given a coarse age partition. All other local authorities in Scotland are labelled "IgLA" and are given a finer age partition. The age by sex partitions used in this calibration group are listed below).

1 = "IgLA Male 0-15"	100 = "smLA Age < 16"
2 = "IgLA Male 16-24"	200 = "smLA Male age 16-39"
3 = "IgLA Male 25-34"	300 = "smLA Male age 40-64"
4 = "IgLA Male 35-44"	400 = "smLA Male age 65+"
5 = "IgLA Male 45-54"	500 = "smLA Female age 16-39"

⁹ Nomenclature of Territory Units (NUTS)

6 = "IgLA Male 55-64"	600 = "smLA Female age 40-59"
7 = "IgLA Male 65-74"	700 = "smLA Female and age 60+"
8 = "IgLA Male or Female 75+"	
9 = "IgLA Female 0-15"	
10 = "IgLA Female 16-24"	
11 = "IgLA Female 25-34"	
12 = "IgLA Female 35-44"	
13 = "IgLA Female 45-59"	
14 = "IgLA Female 60-74"	

Calibration controls used for Northern Ireland

Calibration Group 1: Sex **by** Age (17 group)

Sex (the gender of the respondent)

Age (a 17-category age grouping)

4 = "16-19"	6 = "25-29"	10 = "45-49"	14 = "65-69"
5 = "20-24"	7 = "30-34"	11 = "50-54"	15 = "70-74"
	8 = "35-39"	12 = "55-59"	16 = "75-79"
	9 = "40-44"	13 = "60-64"	17 = "80+"

Calibration Group 2: Local Authority (local authorities in Northern Ireland).

8.3 APPLYING THE WEIGHTS ON THE DATASETS

8.3.1 USING THE HOUSEHOLD WEIGHT FOR GENERAL ANALYSIS

To turn on household weights in SPSS select Data then Weight Cases, and in the dialog box choose to weight cases by the variable HHWT09n (Note n changes each quarter i.e. HHWT092 for April – March 2010 dataset).

Switching the household weight on will change results from numbers of cases interviewed to proportions of the population, so that the total number of cases in the dataset will be equal to the estimated UK population total for the mid-point time period.

The household weight is identical for all members of a household and is calculated as described elsewhere in this section. The household weight is always greater than zero, so each case contributes to the total population.

8.3.2 USING THE SEXUAL IDENTITY WEIGHT FOR SEXUAL IDENTITY ANALYSIS

To turn on the sexual identity weight in SPSS select Data then Weight Cases, and in the dialog box choose to weight cases by the variable SIWT09n (Note n changes each quarter i.e. SIWT092 for April – March 2010 dataset).

The sexual identity weight (SIWT092) is provided for solely for analysis of the Sexual Identity question on the IHS.

The sexual identity weight is calculated for each individual, and is zero for cases when respondents were aged under 16 or were not present in person. Cases with weights of zero will not count towards analysis of results, so the overall population total with the sexual identity weight

enabled will be equal to the adult UK population (i.e. aged 16 and over), with weights for adult non-proxy respondents being correspondingly higher to account for proxy respondents.

When running SPSS datasets weighted by *SIWT09n* a warning message will be generated as SPSS has encountered these zero weights – although this will not affect the results of any analysis. In effect SPSS is filtering the dataset for values of *SIWT09n* greater than zero.

For the Apr09-Mar10 dataset, it would be best practice to filter the dataset for *AGES* greater than or equal to 16 and *PROXY* equal 1 before turning on the non-proxy weight and running analyses, after which the warning message will not appear. For the Jul09-Jun10 dataset and future datasets, filter the dataset for ages equal to 16 and *SIDPROXY* equal to 1, when analysing the sexual identity question.

SECTION 9 – IHS DISSEMINATION AND PUBLICATION

9.1 AVAILABILITY OF ELECTRONIC DATA

9.1.1 SERVICES AVAILABLE FROM ONS

The ONS Customer Response Team can provide tabulations from the IHS. Tables can be provided in Excel.

☎ 01633 455678

✉ socialsurveys@ons.gov.uk

9.1.2 THE DATA ARCHIVE

The UK Data Archive is based at the University of Essex, and holds copies of the End User Licence (EUL) datasets for the IHS. The End Licence is a dataset that has had all disclosive data removed. If Researchers may want to analyse data that is not in the EUL file, they can apply for more detailed data through the Approved Researcher application process. More information can be found at on the ONS website: www.ons.gov.uk/about/who-we-are/our-services/unpublished-data/access-to-ons-data-service

☎ 01206 872001

✉ info@data-archive.ac.uk

🌐 www.data-archive.ac.uk

9.2 PUBLICATIONS

9.2.1 INTERGRATED HOUSEHOLD SURVEY STATISTICAL BULLETIN 2010 – EXPERIMENTAL

In September 2010 a small statistical bulletin will be published alongside the first IHS dataset. The Statistical Bulletin provides a high-level overview of a selection of themes from the survey. For 2010 the Bulletin is based on the Social Wellbeing theme and contains experimental statistics on Sexual Identity, Religion, Ethnicity and Perceived General Health.

☎ 0845 601 3034

🌐 www.statistics.gov.uk/statbase/Product.asp?vlnk=15381

9.2.2 MEASURING SEXUAL IDENTITY: EVALUATION REPORTS – SEXUAL IDENTITY 2010

The IHS is the first ONS dataset to provide data on Sexual Identity. In September 2010 an evaluation report on the sexual identity question and data will be release alongside the publication of the first IHS dataset. This evaluation is based on the IHS experimental data, collected between April 2009 - March 2010.

☎ 0845 601 3034

🌐 www.statistics.gov.uk/statbase/Product.asp?vlnk=15381