



**Scottish Household Survey
Statistical Disclosure Control of 2021 Telephone
Survey Microdata
- UK Data Archive**

Version Control

| Version Number | Drafted by | Published / Draft |
|----------------|---|-------------------|
| V0.1 | Hannah Wolfram, SHS Assistant Statistician, Communities Analytical Division | Draft |
| | | |

Document Approval

| Approved by | Role | Date |
|--------------|--|------------|
| Alison Stout | SHS Project Manager, Communities Analytical Division | 04/10/2023 |
| | | |

1. Background

1.1 The Scottish Household Survey (SHS) is a continuous survey based on a sample of the general population in private residences in Scotland. The survey started in 1999 and up to 2011 followed a fairly consistent survey design. From 2012 onwards, the survey was substantially redesigned to include elements of the Scottish House Condition Survey (SHCS) including the follow-up Physical Survey component. The survey is run through a consortium led by Ipsos MORI.

1.2 Typically, SHS respondents are interviewed face-to-face, in their homes. However, in March 2020 the fieldwork approach was altered in response to the Covid-19 pandemic. This resulted in the majority of the 2020 survey fieldwork, and all of the 2021 survey fieldwork, being carried out using telephone interviewing. As with the 2020 results, the results of the 2021 SHS telephone survey are not directly comparable to SHS face-to-face survey results for previous years (2019 and earlier). The results from the 2020 and 2021 telephone surveys are broadly comparable. However, 2020 data was collected in October 2020 and January-March of 2021, while the 2021 data was collected over the course of a whole year, between April 2021 and March 2022. So users should consider potential seasonal effects when making comparison between the two survey years. Further information on the impact of the change in mode is available in the SHS 2020 and 2021 methodology reports¹.

1.3 The SHS is designed to provide reliable and up-to-date information on the composition, characteristics, attitudes and behaviour of private households and individuals, both nationally and at a sub-national level, and to examine the physical condition of Scotland's homes. It covers a wide range of topics to allow links to be made between different policy areas.

1.4 The specific aims of the survey are to:

- Meet central and local Government needs for priority policy relevant data across a broad range of topics (including needs for continuing time-series of data collected by the SHS and SHCS previously);
- Have built in flexibility to respond to different data needs regarding geography and frequency (e.g. to provide some data annually at Local Authority level, and some biennially at national level), and changes to these requirements over time;
- Align with other surveys and data vehicles (in particular the Scottish Health Survey and Scottish Crime and Justice Survey) to allow for Scottish Surveys Core Questions data to be produced;

¹ [Scottish Household Survey 2021: methodology and fieldwork outcomes - gov.scot \(www.gov.scot\)](https://www.gov.scot/publications/scottish-household-survey-2021/methodology-and-fieldwork-outcomes/pages/11.aspx)
[Scottish Household Survey 2020: methodology and impact of change in mode - gov.scot \(www.gov.scot\)](https://www.gov.scot/publications/scottish-household-survey-2020/methodology-and-impact-of-change-in-mode/pages/11.aspx)

- Produce high quality data in accordance with the Code of Practice for Official Statistics² so as to provide data that is suitable for the production of National Statistics publications in a cost effective way;
- To permit disaggregation of information both geographically and in terms of population sub-groups (such as families with children or households in the social rented sector);

1.5 In order to allow the survey to achieve best value from the data collected, a version of the data is made available on the UK Data Archive to allow researchers to carry out their own analysis. The data is released under End User Licence, and therefore the Scottish Government has taken action to protect the data from disclosure.

2. Introduction

2.1 This document sets out the main ways in which the SHS data has been protected for risk of disclosure, or what is known as statistical disclosure control (SDC).

2.2 This includes action taken on the 2021 survey year's data, including:

2.2.1 SHS social survey data; and

2.2.2 Travel Diary data – journey and stage files including distance to school and distance to work datasets.

2.3 The physical survey data (known as the Scottish House Condition Survey module) is out of scope of this document but is available under End User Licence from the UK Data Archive.

2.4 The Government Statistical Service (GSS) Disclosure Control Guidance for Microdata Produced from Social Surveys (October 2014)³ cites principle 5 of the Code of Practice for Official Statistics⁴ for guidance on the publication of microdata under [End User Licence] EUL, requiring that: “official statistics do not reveal the identity of an individual or organisation, or any private information relating to them, taking into account other relevant sources of information.”

2.5 “Information relating to an individual should be considered by a producer of statistics to be ‘private’ if it was provided with the expectation that the information would be kept out of the public domain. Survey pledges provide respondents with assurances that the information they provide will remain confidential.”

² <https://code.statisticsauthority.gov.uk>

³ [Policy for social survey microdata](#)

⁴ It should be noted that this guidance does not refer to the most recent version of the [Code of Practice for Statistics](#). In the most recent version this principle is covered by the [Data governance](#) principle under the [Trustworthiness](#) pillar.

2.6 Furthermore, the guidance states that “arrangements for confidentiality protection are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics”.

2.7 Private information is defined in the National Statistician’s Guidance; Confidentiality of Official Statistics as being information that:

2.7.1 relates to an identifiable legal or natural person, and

2.7.2 is not in the public domain or common knowledge, and

2.7.3 if disclosed would cause them damage, harm or distress

2.8 Previous SHS sweeps up to the two year fieldwork years 2007/2008 received limited disclosure control. The focus of the disclosure control was around a small amount of top-coding and banding of certain identifying variables.

2.9 This document sets out the main areas of protection that have been applied to the 2021 SHS social survey data which is deposited on the UK Data Archive⁵

3. Statistical Disclosure Control Methods

3.1 For the 2012 SHS main social data (including the travel diary module) a three step process was adopted to reduce the risk of disclosive information being released. The **first step** identified and processed key variables to reduce their risk to disclose private information about respondents. The **second step** was to conduct an intruder test on the amended dataset. This produced a set of recommendations which were implemented in **step three** to produce the final data set described in this report. The resulting disclosure control has been applied to the 2021 SHS main social data (including the travel diary module). All new variables appearing in 2021 have been assessed separately.

3.2 Three methods were used to disclosure control the main data set: re-coding, top-coding and variable removal.

3.2.1 Re-coding is used on categorical variables to collapse response categories together into larger groups to hide small number of respondents in some of the rarer groups (e.g. grouping together rarer ethnicities into broad categories).

3.2.2 Top (or bottom) coding is used on numerical variables where values of a variable above or below a specified level are grouped together and labelled as such. For example, age may be top-coded at 86 and reported as ‘86+’.

3.2.3 Variable removal (or suppression) is used to remove sensitive variables (e.g. sexual orientation). In addition, variables used to calculate other summary variables (e.g. household relationships are used to derive household type) can be disclosive if not recoded or suppressed. These can help in identification of respondents in combination with other variables.

⁵ <http://data-archive.ac.uk/>

3.3 The following sections of this report outline the set of SDC approaches applied to variables in the 2021 SHS Main Dataset (including the travel diary module), using each of these three methods.

4. Suppressed Variables

4.1 A number of variables and records have been removed (suppressed) from the EUL datasets.

| Topic | Variable | Notes |
|-------------------------|--|---|
| Admin | <p>Date of interviews, interview stream, agreement to carry out physical survey (streams 1 to 4 only), random adult interview</p> <p>QUARTER FYEAR H_DATE, DAY, MONTH AND YEAR R_DATE, R_DAY, R_MONTH AND R_YEAR STREAM PS1 PS1A HHPERM RANDPER P2SAMEDAY</p> | Interviews are batched up by area to minimise interviewer travel costs. The interview date could be used to identify the respondent's local area. |
| Geography | <p>Scottish Index of Multiple Deprivation (SIMD) 2020 deciles, 15% most deprived vs. rest of Scotland and older versions of SIMD</p> <p>MD12DEC MD12QUIN MD12PC15 MD16DEC MD16QUIN MD16PC15 MD20DEC MD20PC15</p> <p>Six-fold urban rural classification 2020</p> <p>SHS_6CLA</p> | SIMD 2020 quintiles and two-fold urban rural classification 2016 (re-coded) are provided (see section 0, re-coded variables). |
| Household Matrix | <p>Relationships between household members</p> <p>HA3_1_2 TO HA3_10_9</p> <p>OTH1REL TO OTH3REL</p> | The inclusion of these household composition variables can negate the purpose of summary variables and provide a disclosive risk of identifying large households or other uncommon household types from the data set. |

| Topic | Variable | Notes |
|--|---|--|
| | NUMBHH HHSIZE TOTPEEPS | Any household of more than 6 people was removed from the dataset as they are intrinsically disclosive. |
| School Child | School of random school-age child SCHOOL | School of random school-age child is highly disclosive for some households in certain local authorities. |
| Health and Disability | Health conditions of household members HF2A_P1_01 TO HF2A_P10_20 | Health conditions of household members considered sensitive and could identify rare households. |
| Economic Activity and Education | Economic activity, Number of jobs/hours worked, Employment, National Statistics Socio-Economic Status ILO1_1 TO ILO1_6, ILO2 TO ILO4, RAILO1_1 TO RAILO1_6, RAILO2 TO RAILO4, HG18, HIHWRK2, HG22 TO HG28, HH1 TO HH55, HH1_DKREF TO HH54_DKREF, NSSEC EDQUAL01 TO EDQUAL13, OA1INEMP TO OA3INEMP, OA1EARN2 TO OA3EARN6, EARNSUM_OA1 TO EARNSUM_OA3, EARNSUM, EARNINC, EARNINC_OA1 TO EARNINC_OA3 | Variables that are used to generate summary variables which could be used to disclose information about an individual have been removed. Dataset contains a number of summary variables in place (e.g. NSSEC2). |
| Household finances | Component amounts of net annual income from benefits and miscellaneous sources, summary variables for different components of income (describing level of imputation, if any) and amount Income from benefits <ul style="list-style-type: none"> All variables beginning HH56, HH57 or HH60 <u>that relate to adults other than the highest income</u> | Many income variables are collected during interview in order to derive imputed values based on the “total net annual” income. Main social data includes the main income variables available (total net annual income, income from earning, benefits, miscellaneous sources). |

| Topic | Variable | Notes |
|-------|--|---|
| | <p><u>householder and their spouse</u></p> <ul style="list-style-type: none"> • All variables beginning HH61 or HH62 • All variables beginning BENSUM <u>except</u> BENSUM_HIHSP • All variables beginning BENINC <u>except</u> BENINC_HIHSP <p>Income from miscellaneous sources</p> <ul style="list-style-type: none"> • All variables beginning HH67 to HH71 • All variables beginning MSCSUM <u>except</u> MSCSUM_HIHSP • All variables beginning MSCINC <u>except</u> MSCINC_HIHSP <p>Derived variables on earnings from main and other jobs for the HiH and their spouse</p> <p>HINCMINC, SINCMINC, HINCOINC, SINCOINC, HINCMSUM, SINCMSUM, HINCOSUM, SINCOSUM</p> <p>Other derived variables on household finances</p> <p>WKNETINC, WKNETINC_BROAD, BANDINC, BANDINC_BROAD, INCBAND, INCBAND_BROAD, ANNETINC_BROAD, INCSUM_BROAD, TOTHTINC_BROAD</p> <p>Council tax band and total and after deductions council tax payments</p> <p>COUNCILTAXBAND CT2A CT2B</p> | <p>Many summary variables have been included (e.g. BENINC_HIHSP; total net annual income from benefits for the highest income householder and their spouse)</p> |

| Topic | Variable | Notes |
|--------------------------------------|---|--|
| Housing | Mortgage/rent payments HJ2 to HJ6 | These variables have been excluded from the dataset. However, the new derived variables on housing costs (e.g. mortgage_amt and rent_amt) are provided. |
| Transport and Travel | Name of school/college/university, location of workplace RD1 RD2 | Name of school/university and location of the place where you work is highly disclosive for some households in certain local authorities. |
| Composition / Characteristics | Sexual orientation RAND_ID | Deemed sensitive for release of EUL datasets |
| | Reason for discrimination or harassment being 'your sexual orientation' or 'your trans status, including non-binary identities'. DISCRIM_09, DISCRIM_10, DISCRIM_15, HARASS_09, HARASS_10, HARASS_15 | High level of association between reason for discrimination and/or harassment being your sexual orientation and respondent identifying as lesbian/gay. Small number of respondents reporting the reason for discrimination and/or harassment being your trans status, including non-binary identities. |
| | Gender HIHGENDER RANDGENDER SPGENDER OTH1GENDER OTH2GENDER OTH3GENDER GENDER HA6_1 TO HA6_10 | Therefore these have been combined with the 'other' reason discrimination and harassment variables to create new variables DISCRIM_091015 and HARASS_091015 |
| | | One tenth of one per cent (0.1 per cent) of adults described their gender as 'in another way' or did not provide an answer. As re-coding is not appropriate for the gender variables, any households where a respondent has described their gender as 'in another way' or did not provide an answer have been removed. |
| Number of Children | Total number of children and total number of eligible children in the household TOTKIDS NUMKIDS | Households containing 5 or more children have been removed from the data because households with large numbers of children present are easily identifiable. For the SHS 2021, this affects 6 households. |
| Number of Adults | Total number of adults and eligible adults in the household TOTADS NUMADS | Households containing 6 or more adults have been removed from the data because households with large numbers of adults present are easily identifiable. For the SHS 2021 this affects 8 households. |

| Topic | Variable | Notes |
|--------------------------------------|---|---|
| Childcare and childcare costs | Childcare costs and childcare involving overnight stays | |
| | ELC2A ELC2B ELC3A ELC3B ELC11 ELC11A ELC12 ELC12A ELC17_KIDOTH1 TO ELC17_KIDOTH5 ELC17A_KIDOTH1 TO ELC17A_KIDOTH5 ELC18_KIDOTH1 TO ELC18_KIDOTH5 ELC18A_KIDOTH1 TO ELC18A_KIDOTH5 ELC20 ELC20A ELC20B | <p>A small number of households reported childcare arrangements involving overnight stays. The variables allowing these households to be identified have been excluded.</p> <p>Childcare costs are reported as either weekly or monthly. Therefore, there is no easy way of top coding these variables, so they have been excluded. However, the derived annualised childcare costs variables are included and are top coded.</p> |
| Redundant | Council tax instalments | After excluding the council tax band and council payment variables CT2A and CT2B which are disclosive (see housing topic), CT3 becomes redundant |
| | CT3 | |
| | Age of respondents/spouse/random school child | Single year of age provided for highest income householder and random adult. |
| | HIHAGEBD SPAGE AGERANK AGERBAND OTH1AGE TO OTH3AGE AGE_7 TO AGE_10 PENAGE_7 TO PENAGE_10 HA6_7 TO HA6_10 HA7_7 TO HA7_10 HA10_7 TO HA10_10 HA13_7 TO HA13_10 HF1A_07 TO HF1A_10 | <p>Banded age is provided for all other persons.</p> <p>Any household of more than 6 people was removed from the dataset as they are intrinsically disclosive. Therefore the variables AGE, PENAGE, HA6, HA7, HA10, HA13 for persons 7 to 10 become redundant.</p> |
| | NUMCARS | A new variable NUMCARS_NEW is provided which details number of cars available to household (NUMCARS included number of vans as well) |
| | HA6_P1_OTHERWAY TO HA6_P10_OTHERWAY | See the notes on the gender variables in the composition and characteristics topic. |

5. Recoded Variables

5.1 Recoded variables are usually categorical variables. In most cases this involves combining smaller categories to reduce the risk of disclosure of a household adult.

| Topic | Variable | Notes |
|--------------------------------------|--|--|
| Admin | <p>Serial number, survey weights</p> <p>UNIQID LA_WT2 IND_WT2 KID_WT2 TRAV_WT2</p> <p>LA_GRWT2 IND_GRWT2 KID_GRWT2</p> <p>LA_WT3 IND_WT3 KID_WT3</p> <p>LA_GRWT3 IND_GRWT3 KID_GRWT3</p> | <p>The unique identifier (serial) variable is recoded to hide information about the order the interviews were carried out in, as this can be linked to local areas. This recoding is consistent with the travel diary module.</p> <p>Survey weights are rounded to two decimal places.</p> |
| Composition / Characteristics | <p>Religion and Ethnicity</p> <p>hih_rel RANDRELG HIH_STAT2012 RANDSTAT2012 HIH_ETH 2012 RANDETH2012</p> | <p>Religion and ethnicity are key variables that could lead to the identification of small groups of people.</p> <p>Religion has been recoded to group all non-Christian religious groups together, due to the small numbers of respondents in these groups. Ethnicity was recoded to group small white ethnic groups together and non-white minority ethnic groups together.</p> <p>Marital status of highest income householder and random adult are coded to bring together similar groups of marriage and civil partnership (e.g. married / civil partnership, divorced/legally dissolved civil partnership)</p> |

| Topic | Variable | Notes |
|-----------------------------|--|---|
| | Country of birth RANDFROM2012 | Due to small base sizes for UK countries (excluding Scotland) and the Republic of Ireland as the birth country of the respondent, the country of birth is a key variable which could lead to identification of respondents. The UK countries excluding Scotland have been grouped together and the Republic of Ireland and the Elsewhere category have also been grouped together. This approach is consistent with the pooled sample. |
| | AGE_1 to AGE_6 | Ages of individuals in the household have been banded by rough school age groups (0-4, 5-11, 12-15) and then five year age bands. |
| Geographic | Six-fold urban rural classification 2016 and Scottish Index of Multiple Deprivation (SIMD) 2020 quintiles SHS_6CLA MD20QUIN | SHS_6CLA (six-fold urban rural classification 2016) has been recoded to the two-fold (urban or rural) classification (SHS_2CLA) as it could disclose unique data zones of respondents. A small number of data zones can be uniquely identified by combinations of council, SIMD 2020 quintile and two-fold urban rural classification 2016. In these cases the urban rural classification was changed. Generally this involved changing the classification from rural to urban. If this resulted in an implausible combination then the SIMD quintile was changed. |
| Housing | Accommodation type HB1 Rent payments RH4 and RH8 | Some rare types of accommodation have been coded together. Questions on the number of people in the household contributing to rent payments and on what other costs rent payments included have been recoded. |
| Transport and Travel | Main mode of travel (in the travel diary dataset) MAINMODE | Smaller categories of mode of travel have been combined |
| | Two-fold urban/rural classification for journey/stage start/end points (in the travel diary dataset) JOURNEYSTART_RURURB JOURNEYEND_RURURB STAGESTART_RURURB STAGEEND_RURURB | Unique combinations of LA/ Urban rural have been re-coded for consistency with main social data. |

| Topic | Variable | Notes |
|---------------------------------------|--|---|
| Economic Activity / Employment | Socio-economic variables HIHSOC HIHSIC | Banded together into broad analytical classes. |
| Young People | Local Authority of the school the children go to SCLCOUN Laschool (travel diary) | Recoded to group together where the local authority of a school is different to the household local authority, as this could lead to the identification of rare households. |
| Childcare and childcare costs | Hours of unfunded/funded childcare per week ELC2 ELC3 ELC8 ELC4D | Recoded into the following groups to reduce the risk of identifying respondents with unique childcare hours. Don't know Refused None/Less than 1 hour 1-10 hours 11-20 hours 21-30 hours Over 30 hours Don't know Refused 1-10 hours 11-20 hours 21-30 hours over 30 hours |

6. Top (or Bottom) Coding Variables

6.1 Many numerical variables may identify households/adults due to values that are unusual or very low/high (e.g. very young/old adults).

| Topic | Variable | Notes |
|--|---|---|
| Household Composition / Characteristics | Age of highest income household and random adult HIHAGE RANDAGE | Top-coded at 86+. HIHAGE also bottom coded at '18 and under' |
| | Year of arrival in UK MIGBIR2_Y | Reference to very old years may identify people who arrived in the UK a very long time. MIGBIR2_Y is top-coded at 30+ years. |
| Housing | Date first started renting/own property | Reference to very old years may identify people that have lived in areas for a very long time. |

| Topic | Variable | Notes |
|--------------------------------------|---|--|
| | T38YEAR T39YEAR RA1 | T38YEAR, T39YEAR, and RA1 are top-coded at 30+ years. |
| | Number of bedrooms HC4 | May lead to identification of very large, rare, households. Top-coded at 5+. |
| | Number of housing lists on RA15A | May lead to the identification of rare households. Top-coded at 2+ |
| | Purchase price/Amount borrowed and housing costs OH9 OH14 OH15 OH16 MORTGAGE_AMT RENT_AMT SHARED_OWNERSHIP_AMT HCOST_AMT | May lead to identification of very expensive households or households with very high costs. Top-coded at 95 th percentile and bottom-coded at the 5 th percentile. |
| Transport and Travel | Number of cars, amount spent on fuel, number of bicycles NUMCARS_NEW NUMVANS HD19 HD2011 PARKCOST PTCOST | May lead to identification of households with very large number of cars, offers limited utility. NUMCARS_NEW top-coded at 2+ and numvans top-coded at 1+. Amount spent on fuel very high for some individuals. Top-coded at 95 th percentile. Amount spent on parking and public transport top-coded at 95 th percentile. Number of bicycles top-coded at 3+. |
| Household Finances | Income from benefits, miscellaneous sources and earnings. BENINC_HIHSP MSCINC_HIHSP EARNINC_HIHSP ANNETINC | Very large incomes should be protected as it may lead to identification of households and the people within them. All income variables are top-coded at the respective 95 th percentile. |
| Childcare and childcare costs | CCCOST_COST RANELCKID_COSTS_TERM RANELCKID_COSTS_HOL OTHCCCKID1_COSTS_TERM OTHCCCKID1_COSTS_HOL OTHCCCKID2_COSTS_TERM OTHCCCKID2_COSTS_HOL OTHCCCKID3_COSTS_TERM | All annual childcare costs variables top-coded at the respective 95 th percentile. |

| Topic | Variable | Notes |
|-------|----------------------|-------|
| | OTHCCCKID3_COSTS_HOL | |