

Scotland's Census 2022

Coding Methodology

September 2021

Contents

1. Introduction.....	3
1.1 Purpose of this document	3
2. Data capture and coding.....	3
2.1 Overview and approach	3
2.2 Coding specification.....	4
3. Online coding.....	5
3.1 Online coding overview.....	5
3.2 Online system features for capture and coding.....	5
3.3 Testing of the online system	6
4. Paper coding.....	7
4.1 Paper coding of census and census coverage survey questionnaires	7
4.2 Examples of paper coding methods.....	7
4.3 Address coding methods	8
4.4 Testing of paper coding	9
5. Manual coding.....	10
6. Lessons learnt from the October 2019 Rehearsal	10
7. Improvements since the 2011 Census.....	11
8. Harmonisation with Office for National Statistics.....	11
10. Appendix A.....	13

1. Introduction

1.1 Purpose of this document

This paper outlines the methods and processes National Records of Scotland (NRS) have set up for the capture and coding of Scotland's Census 2022 and Census Coverage Survey (CCS) questionnaires. The paper provides an overview of the approach taken, outlines the differing methods for online and paper coding and highlights the improvements to data quality since 2011. A summary of the extent to which coding methods harmonise with other UK countries is also provided.

2. Data capture and coding

2.1 Overview and approach

The 2022 census will be the first one conducted primarily online. A key benefit of an online platform is that by improving the user experience, and improving efficiencies in collection and processing of information, data quality will be maximised. Those who wish to complete a paper questionnaire will still be able to do so as it is important to achieve high response rates for all groups of the population. It will only be possible to complete the CCS via paper questionnaire although the majority of these will be filled in by an interviewer rather than a member of the public.

Responses to census questions are given numeric codes in order to process data and produce outputs. This is referred to as coding. We have adopted a number of principles when developing our census coding methods to ensure data quality standards are achieved. These principles include:

- Coding at the point of collection, as the data is captured, through “**autocoding**” techniques. Autocoding means that the census coding logic is embedded into the Online Collection Instrument (OCI) and the software developed to scan and code paper questionnaires.
- Ensuring responses are given codes from consistent, recognised classifications irrespective of whether submitted via paper or online. The way in which codes are assigned is governed by **coding specifications**, which have been developed and managed by the NRS Census Coding Team.

- **Reducing the dependence on manually correcting** or inspecting responses. The added benefit of setting up autocoding techniques at source will support this and reduce human error in manually correcting records.
- To **harmonise** with Office for National Statistics and NISRA where possible.

2.2 Coding specification

The coding logic that details how to code census question responses is contained in a document called the **census coding specification**. It is also a key reference document used by others for statistical processing and analysis. There is a separate coding specification for the CCS questionnaires. The coding specifications contain a number of elements including:

- Instructions for IT systems to align to, known as **business rules**. Appendix A includes some examples of these for tick, text and address questions.
- **Classification indexes**, which provide the listing we want responses to code to. These indexes define a one to one relationship between a numeric code and its categorical label. These labels will be the basis for published outputs.
- **Coding indexes**, which contain lists of synonyms or alternative responses that code back to the classification indexes. These indexes define a many to one relationship between a numeric code and text responses. An example of this would include a respondent stating “US of A” as their country of birth but the coding index would instruct the response to be coded as “United States” (see Table 1 below).

Coding index		Classification index	
THE UNITED STATES OF AMERICA	840	JERSEY	832
U S	840	ISLE OF MAN	833
U S A	840	TANZANIA	834
U S OF A	840	UNITED STATES	840
UN STATES OF AMERICA	840	UNITED STATES	
UNITED S AMERICA	840	VIRGIN ISLANDS	850
UNITED S O AMER	840	BURKINA	854
UNITED ST OF AMERICA	840	URUGUAY	858
UNITED STATES	840	UZBEKISTAN	860

UNITED STATES AMERICA	840
UNITED STATES OF AMERICA	840
UNITED STATES USA	840
UNITED USA	840
UNITEDSTATESAMERICA	840
UNITEDSTATESOFAMERICA	840
US	840
USA	840

Table 1 An example of coding index and classification index for Country of Birth.

To improve on data quality, the census coding specification was refreshed following lessons learnt from the October 2019 rehearsal evaluation. This included updates for incorrect business rules and adding entries to the coding indexes. The work invested in expanding the coding indexes will increase the match rate and reduce the need to manually inspect records during live operations.

3. Online coding

3.1 Online coding overview

Scotland's Census 2011 was the first time people in Scotland were able to respond to their census using an online platform. The 2022 census is designed to maximise the number of responses collected online through the Online Collection Instrument (OCI).

3.2 Online system features for capture and coding

For responses collected online, the quality of data is high. In addition to complying with the coding specification (section 2.2) there are a number of features that have been incorporated into the OCI that ensure consistency of coded responses, including:

- Only showing respondents the questions they need to answer ("**routing**"). Examples of this would be that those under the age of 16 are not shown the voluntary sexual orientation question or those who have never worked will not be asked about their current or most recent occupation or industry.
- **Validating responses** in real time which helps to keep data clean by, for example, only allowing expected character types for certain questions (for

example, numeric characters for questions that ask for dates or letter characters for questions that ask for text responses).

- Special features such as **radio buttons**, which only allow a single response to be selected for single tick questions,
- **User-optimised indexes** provide a list of finite suggested options to select an answer from. For example, users will be prompted to answer from a drop down list of countries for the “country of birth” question. When an option is picked from the list the corresponding code is stored. This feature avoids respondents entering misspelt or erroneous answers.
- Real-time validation of responses also ensures that mandatory questions are answered and pop-up reminders can be triggered in instances of non-response to questions. Similarly, validations are also in place to ensure that there is a consistent logic between questions (e.g. ensuring that a respondent’s marital status is consistent with their stated relationship to others in the household).

While routing and validation were applied to the 2011 online questionnaire, type-ahead and address look-up functionality for text based and address questions have been added for 2022. Type-ahead functionality allows the respondent to enter text and choose from a list of options which match the text entered. This is enhanced for address questions, which allow the respondent to search for their address by typing in either their postcode or the first line of their address. These features allow further improvement of the internal consistency and quality of the data.

3.3 Testing of the online system

The online system has gone through a number of rigorous test cycles to ensure the functionality set out above produces data coded in the way required. It has been tested to simulate a variety of households completing the form (e.g. single person, families, co-habiting couples, student households etc). This was done to ensure that the validation suite, user optimised lists and the routing work as expected for the different household types in Scotland.

4. Paper coding

4.1 Paper coding of census and census coverage survey questionnaires

Paper questionnaires are scanned and the responses are captured digitally using character recognition software. Simple tick based questions are coded following the rules of the coding specification (Appendix A). Additional techniques are applied to code text based questions to correct for things like spelling mistakes and word ordering (see section 4.2).

Coding accuracy and completeness is lower for paper questionnaires than for online questionnaires because there are limits to interpreting handwriting. In addition, the completion of paper questionnaires does not benefit from the checks that the online validation and routing provide.

4.2 Examples of paper coding methods

For the Scotland's Census 2022 additional coding techniques will be applied to text and address questions to help assign a code when the response does not exactly match the coding index. These are presented in the table below.

Text coding techniques for paper questionnaires

Technique	Description
N-Grams	<p>This accounts for variation in word ordering of responses. Text is matched against the sequence of words (N-Grams) in the response from the longest to shortest order. If a match isn't found against the coding index to the longest N-Gram, matches are attempted against the next shortest N-Gram from left to right.</p> <p>For example, Seoul South Korea can be broken into:</p> <ul style="list-style-type: none"> • three word n-gram (Seoul South Korea), • two word n-grams (Seoul South and South Korea) • single word n-grams (Seoul, South and Korea)
Distance matching	<p>Allows insertions, deletions or substitutions of characters so that the response text exactly matches an entry in the coding index. This can help provide a coded match when a word has a minor spelling mistakes. The length of a response determines number of changes allowed (longer responses are allowed more changes).</p> <p>For example:</p>

	<ul style="list-style-type: none"> • “FRANCF” is one change away from “FRANCE” • “AUSTRAILA” is two changes away from “AUSTRALIA”
Stemming (only used in some text questions)	<p>This reduces words to their ‘stem’ or root definition.</p> <p>For example for the employer activity (industry) question, the words “CONSULTANT” and “CONSULTING” are both stemmed to “CONSULT”. The coding index is also stemmed to match against.</p>
Dictionary matching (only used in some text questions)	<p>This is where a match is found using only part of a phrase from the coding index provided.</p> <p>For example for long term health conditions if the response is “PROBLEM WITH VOCAL CHORDS” then this allows us to match to “VOCAL CHORDS INJURY” as it matches the key words “VOCAL CHORDS”.</p>

These additional techniques will result in higher rates of automated coding than those achieved in previous years. In 2011 the approach was to match responses directly against the coding index only with no accounting for commonality in words (such as with the stemming) or word ordering (N-Grams). The distance matching technique will realise a much higher match rate against the coding index as accounts for spelling mistakes.

4.3 Address coding methods

The aim of coding addresses is to return a valid postcode but this can be difficult from paper questionnaires as often the postcode is not provided or the complete address is not given. Additionally the similarity of addresses, particularly street names across the UK, meant that additional techniques (or processing steps) were required to code an accurate postcode. These were developed as follows:

- **Abbreviation translation:** This changes common abbreviations in address’ to allow for exact matching, for example, Rd = Road.
- **Town/city matching:** This is done to identify which town/city the address is from as street and road names tend to not be very unique. Distance matching of the town/city occurs at this step.
- **Address Cracking:** This splits the address into the various components like building name, building number, road, town/city.

- **Address Match:** If possible, subset of address created using town/city and house number. A scoring mechanism is then used which considers things like:
 - Words which exactly match.
 - Words that are distance matched.
 - Number of words that matched compared to number of unmatched word

If adequate matching is obtained for the address components listed above it is more likely that valid coded address can be returned.

4.4 Testing of paper coding

Given the limitations for paper questionnaires, NRS staff worked closely with the suppliers during development and put in place a rigorous test plan to test the different techniques. Part of this testing involved NRS staff using responses from the Scotland's Census 2011 to test the accuracy of the coded values. The 2011 responses are thought to be indicative of how people may fill in their paper forms in 2022 and provides common misspellings or shortening of words that we would like to code. For questions that were new to Scotland's Census 2022, test cases had to be substituted from another question, for example the new question about passports used data from the 2011 country of birth and national identity questions.

As address questions are often poorly completed, real addresses from the UK address list were used with some altered to include spelling mistakes and remove parts of the address like the postcode. In adjusting the responses in this way a full range of address data quality could be assessed and we were able to fully test the limits of the address matching tool.

Additionally, configuration testing was undertaken to gauge the optimum settings for coding text responses. For example, we tested the impact of changing the number of letter differences a word could be misspelt by for the distance matching technique. Another example of configuration testing was to switch the stemming technique off and on. An assessment was made to try to achieve the best balance of accuracy of coding against the number of responses coded, in order to reduce the burden on manual coding.

5. Manual coding

Whilst new techniques have been developed to increase the coding rate and accuracy of Scotland's Census 2022 responses via autocoding, there will be a proportion of answers that will need to be manually inspected by a person to assign a code. When the census is running a "manual coding operation" will be stood up. The aim of the manual coding operation is to fill in the gaps and help provide as complete a dataset as possible.

The manual coding operation will be resourced by a number of temporary "coders" and "supervisors". They will be trained to understand the coding indexes and will assign valid codes where the automated techniques have been unable to do so.

It is anticipated that manual coding will be required more so for responses collected on paper forms. In these cases coders will consider poor handwriting the scanning software cannot clearly interpret. In addition, it is expected that a relatively high number of occupation, industry and address questions will require manual coding due to variations in natural language when describing jobs, variations in address formats, and the large number of options these responses can be coded to.

Where the contracted coders cannot allocate a valid code, there will be processes in place to escalate to NRS subject experts. They will identify a valid code if possible and provide feedback to improve the knowledge base for coders going forward. It won't always be possible to identify a code, particularly if limited information has been provided on paper forms. These responses that remain as un-coded will be dealt with further downstream at the [Edit and Imputation](#) stage of data processing.

6. Lessons learnt from the October 2019 Rehearsal

A rehearsal of the census was undertaken in October 2019 with responses from 18,500 households allowing the opportunity to establish how effectively questions were coded in order to inform improvements for the live census.

Analysis from the rehearsal informed improvements to our coding techniques. Examples of improvements included updating our lists of expected text responses as well as influencing the treatment of blank and partial census responses. These changes were detailed in our Coding Specification document and formed the basis of subsequent development by our contractors. Analysis of the rehearsal data also

informed changes to how questions answered online are validated and how some users are routed through to only see questions relevant to them. More information can be found in the [rehearsal report](#).

7. Improvements since Scotland's Census 2011

In summary, the main improvements to coding of Scotland's Census 2022 results are:

- A higher quality of data accuracy and completeness as it will be carried out primarily online with the added benefits of question routing, validation and other online functionality which codes the data at source (section 3.1). These techniques are more sophisticated by comparison to the online data collected in 2011.
- The quality of paper coded questionnaire responses will also be higher due to the investment in the new paper coding methods (section 4). These types of methods were not developed in 2011 and the coding rates were considerably lower and of poorer accuracy.
- Due to the investment in the automated techniques for both online and paper, the need for manual coding (i.e. human intervention) is much lower in 2022.

8. Harmonisation with Office for National Statistics

NRS have worked with other UK census departments to ensure a level of harmonisation is achieved. With respect to coding, the logic and indexes used to code census responses have been developed in consultation with the other UK statistical agencies, the Office for National Statistics (ONS) and the Northern Ireland Statistical and Research Agency (NISRA).

Examples where we have harmonised include liaising with ONS when constructing coding lists and indexes to help fill any gaps or identify inconsistencies when comparing their lists with ours. Discussions at 4-nations meetings ensured agreement on harmonisation on issues such as question routing (or divergence, e.g. in the case of marriage ages differing in Scotland and England). We also liaised with

ONS on the construction of the online user optimised lists to maintain commonality where possible.

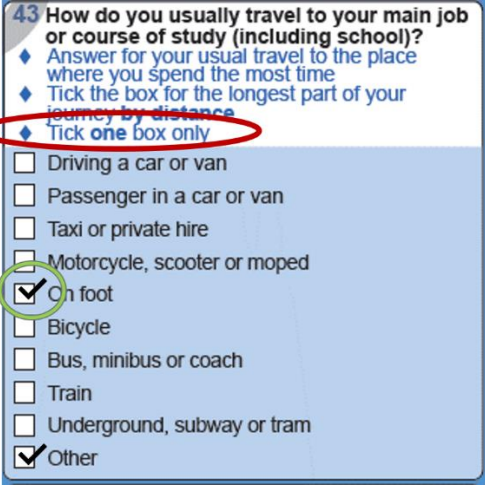
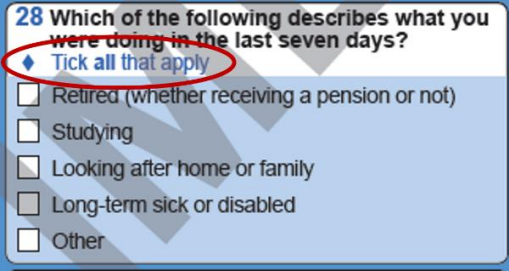
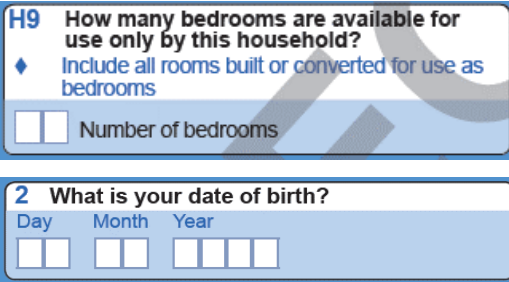
For some questions, however, it is not possible to completely harmonise. This is due to differences in question wording, definitions, response categories or question structure and stakeholder needs. Examples of differences include:

- **Long term health:** Scotland has a write-in box for this question whereas ONS do not have this option. This will feed through to how the coding logic is set up and the variety of responses we will receive.
- **Religion:** Different question wording is used (“What is your religion?” from ONS vs “What religion, religious denomination or body do you belong to?” from NRS).
- **Occupation and Industry:** NRS have implemented drop down lists for users to pick responses from in the online questionnaire whereas ONS have not. This will improve the response completeness for this question but there may be biases for Scottish data as cut down versions of the industry and occupation classification lists are presented. Cut down versions were used as presenting the whole listing for occupation (circa 30,000 records) and industry (circa 16,000 records) would have impacted greatly on system performance.

Through liaising with ONS we have established that some of the auto coding techniques are similar although because they have been developed separately they wouldn't necessarily produce exactly the same results. This would cover techniques like distance matching to correct for spelling mistakes.

Appendix A

A list of the main question types from the census question set and an example of the coding logic used to assign responses numeric codes.

Question type	Coding logic
<p>1. Single tick response</p> 	<ul style="list-style-type: none"> • If a single tick response is captured, this is easy to code automatically. • Some multiple tick responses are resolved in the coding specification but all other instances are flagged for further investigation. • Radio buttons are included in the online questionnaire so only one answer can be selected.
<p>2. Multiple tick response</p> 	<ul style="list-style-type: none"> • Many multiple tick responses can be coded automatically. • Some multiple tick responses are resolved in the coding specification but all other instances are flagged for further investigation.
<p>3. Number</p> 	<ul style="list-style-type: none"> • Numeric responses can be coded automatically. • Some built-in checks included in the coding logic: <ul style="list-style-type: none"> ○ Validating a specific number of digits ○ Converting word equivalents to numbers ○ Validating against alphabetic characters online ○ Converting American date format (MMDDYYYY) to standard date format (DDMMYYYY)

<p>16 What is your main language? ◆ Tick one box only</p> <p><input checked="" type="checkbox"/> English</p> <p><input type="checkbox"/> Other, please write in (including BSL and TACTILE BSL):</p> <p>Weegie</p>	<ul style="list-style-type: none"> Another text example where interpretation by a manual coder is required: <table border="1"> <thead> <tr> <th>Text response</th> <th>Code</th> </tr> </thead> <tbody> <tr><td>Gaelic Scottish</td><td>61005</td></tr> <tr><td>Lowland Scots</td><td>61107</td></tr> <tr><td>Scot Sign Lang</td><td>69801</td></tr> <tr><td>Scots</td><td>61107</td></tr> <tr><td>Scots Gaelic</td><td>61005</td></tr> <tr><td>Scottish</td><td>61107</td></tr> <tr><td>Scottish Cant</td><td>69503</td></tr> <tr><td>Scottish English</td><td>44201</td></tr> <tr><td>Scottish Sign Language</td><td>69801</td></tr> <tr><td>Scottish Standard English</td><td>44201</td></tr> <tr><td>Scottish Traveller Cant</td><td>69503</td></tr> <tr><td>Traveller Scottish</td><td>69503</td></tr> <tr><td>Ulster Scots</td><td>61107</td></tr> </tbody> </table>	Text response	Code	Gaelic Scottish	61005	Lowland Scots	61107	Scot Sign Lang	69801	Scots	61107	Scots Gaelic	61005	Scottish	61107	Scottish Cant	69503	Scottish English	44201	Scottish Sign Language	69801	Scottish Standard English	44201	Scottish Traveller Cant	69503	Traveller Scottish	69503	Ulster Scots	61107
Text response	Code																												
Gaelic Scottish	61005																												
Lowland Scots	61107																												
Scot Sign Lang	69801																												
Scots	61107																												
Scots Gaelic	61005																												
Scottish	61107																												
Scottish Cant	69503																												
Scottish English	44201																												
Scottish Sign Language	69801																												
Scottish Standard English	44201																												
Scottish Traveller Cant	69503																												
Traveller Scottish	69503																												
Ulster Scots	61107																												
<p>5. Combinations of previous types</p> <p>H14 In total, how many cars or vans are owned, or are available for use, by members of this household? ◆ Include any company car(s) or van(s) available for private use</p> <p><input type="checkbox"/> None</p> <p><input type="checkbox"/> 1</p> <p><input checked="" type="checkbox"/> 2</p> <p><input type="checkbox"/> 3</p> <p><input checked="" type="checkbox"/> 4 or more, please write in number 7</p>	<ul style="list-style-type: none"> For questions that combine Tick and Number or Text elements, individual parts are coded according to their separate coding logic and then the coded elements are combined based on priority rules. In general, any written response is prioritised over tick responses. 																												
<p>6. Address</p> <p>11 One year ago, what was your usual address? ◆ If you had no usual address one year ago, state the address where you were staying</p> <p><input type="checkbox"/> Same as Person 1</p> <p><input type="checkbox"/> The address on the front of the questionnaire</p> <p><input type="checkbox"/> Student term-time / boarding school address in the UK, please write in below:</p> <p><input type="checkbox"/> Another address in the UK, please write in:</p> <p>Postcode</p> <p><input type="checkbox"/> Outside the UK, please write in country:</p>	<ul style="list-style-type: none"> Special consideration is given to address questions: <ul style="list-style-type: none"> Address and post code finders are used for online responses Written responses are captured as one long string Various techniques are used to break this string down differently. A successfully coded address is one where a valid postcode is assigned. 																												