Food and You 2: Technical Report: Data validation and management

Overview

Questionnaire versions

As described in earlier sections, the data are collected from two sources: an online questionnaire and either one or two postal questionnaires. The online questionnaire includes some built-in routing and checks within it, whereas the postal questionnaires rely on correct navigation by participants and there is no constraint on the answers they can give.

In addition, the online data are available immediately in their raw form, however the postal questionnaire data must be scanned and keyed (typed in) as part of a separate process. Tick box answers are captured by scanning, and numbers and other verbatim answers are captured by keying, with the data then coded in an ascii text string.

In line with standard procedures on a mixed-mode survey such as this, the online questionnaire is taken as the basis for data processing. Once that is processed then a data map/dictionary is used to match the data from the postal questionnaires with the online data.

A wide range of edits are carried out on the data followed by numerous checks. These have been detailed throughout this section.

Data editing

Postal data – forced edits

The postal data are subject to errors introduced by participants. Edits are required for this data and in some cases to also match the postal routing to the online questionnaire routing. There are five key principles to editing postal data which are drawn upon for this:

1. Forward editing is applied to all filtered questions. If a participant is eligible to answer a question but has not, they were assigned a code of -99 “Not stated”.
2. A small number of back edits are applied to a handful of variables. If a participant has answered a question but has not answered “yes” at the previous filter question a back edit is applied (i.e. the original question is edited so that the data matches the routing). This is only done on variables specified by the FSA as the forward editing approach handles the majority of the cleaning required.
3. A specification is agreed with the FSA that sets out a number of variables which need to be edited to directly match the online routing. This is applied as a post field edit to the postal data only.
4. If a question is incorrectly answered as a multi-code question then the responses are set to -99 “Not stated”.

5. On a handful of questions that offer a multi-code answer we ask participants to limit their answers to a maximum of three, so edits are made where additional answers are given. A random selection is made of the given answers in SPSS and the process ensures no duplicate answer can be selected.

In addition to this, where there is a multi-code variable that also has an exclusive code (such as “don’t know”), answers are edited so that valid multi-code options take priority, and conflicting exclusive codes are deleted. Where there are several exclusive codes, a hierarchy is applied.

Edits to numeric answers

In Wave 6, edits were only made to one question where the answer was deemed to be improbable or unlikely. For ‘Number of adults’, if a participant from a multiple response household answered that only one adult lived in that household a post-field edit was applied to set the answer to two. This edit will have a subsequent impact on any variables that use nadult as part of the filter and therefore some questions will highlight a group that look eligible to answer but did not.

In Waves 1-4, it has also been necessary to edit ‘age’ data. However, due to the inclusion of both the open question recording age in years and the question capturing the same information via pre-defined categories on the postal questionnaires in Waves 5-7, it was possible to use the answers to the age bands question to verify the answers to the open question in the postal data.

Duplicate responses

Duplicate responses are received each wave where participants complete the postal version of the questionnaire as well as the online. The number of duplicate responses for each wave can be found in the accompanying technical report tables. The online version takes precedent, and the postal version is deleted.

Break off rates

The number of break offs (i.e. where a participant has started the online version but abandoned it) and the variable they occurred for in each wave can be found in the accompanying technical report tables.

Coding

Coding is done by Ipsos on one open ended question (FOODISSA2). Coding is the process of analysing the content of each response based on a system where unique summary ‘codes’ are applied to specific words or phrases contained in the text of the response. The application of these summary codes and sub-codes to the content of the responses allows systematic analysis of the data.

Translation of verbatims in Welsh

Participants are able to complete the survey in English and in Welsh. There are a small number of participants who choose to complete the survey in Welsh and provide verbatim text. These verbatims are translated by the FSA’s Welsh Language Unit before being coded, alongside the English responses, by Ipsos.

Ipsos coding

Having established the codeframe for FOODISSA2 “What are your concerns about the food you eat?” in Wave 1 (using Q.1a. “What food issues, if any, are you concerned about?” from Wave 17 of the FSA’s Public Attitudes Tracker as a basis for the codeframe) this coding framework is then updated throughout the analysis process of every wave to ensure that any newly emerging themes are captured. Developing the coding framework in this way ensures that it provides an accurate representation of what participants say. After adding in any new codes to the codeframe, it is then reviewed by the FSA and Ipsos research teams with queries subsequently addressed by the coding team. After this it is then appended to the datasets.

Codes are grouped together into broad themes (e.g. ‘Environmental and Ethical Concerns’), shown in bold text in the data tables. Some of the broad themes also have sub-themes (e.g. ‘Fair Trade / Ethical’). For consistency between waves, all codes developed for previous codeframes are included in the codeframe for the latest wave, including codes for which no responses were assigned. These codes are also present in the data tables (and are marked as having received no responses).

Ipsos uses a web-based system called Ascribe to manage the coding of all the text in the responses. Ascribe is a system which has been used on numerous large-scale projects. Responses are uploaded into the Ascribe system, where members of the Ipsos coding team then work systematically through the comments and apply a code to each relevant piece of text.

The Ascribe system allows for detailed monitoring of coding progress, and the organic development of the coding framework (i.e. the addition of new codes to new comments). A team of coders work to review all the responses after they are uploaded on Ascribe, with checks carried out on 5% of responses.

Data checks

Checks on data

Ipsos check the data in two ways. Firstly, the data is checked using the questionnaire and applying a check for each filter to ascertain whether a participant correctly followed the routing. This checks 100% of the questionnaire and is run separately on the raw postal data and the raw online data. Once the data is checked a list is produced that identifies which variables require an edit and this largely relates to the postal data. Any edits applied are set out in the section ‘Data editing’.

Once the data edits are applied a combined dataset is created, duplicate participants are removed (as outlined in the section on duplicate responses) and then the derived variables are created.

Checks on derived variables

Derived variables are created in syntax and are based on the table specification. All derived variables are checked against previous waves to ensure the values are roughly in line with what we would expect to see. Cross checks are carried out on the syntax used to create the derivations to ensure the logic is valid.

Once the derivations are set up the dataset is checked by other members of the team. Some derived variables are based on one question (for instance age), and these are checked by running tabulations on SPSS from the question they are derived from, to check that the codes feed into the groups on the cross-breaks. If the derived variables are more complex and based on more than one question, e.g. NS-SEC, more thorough checks are carried out. For example, the NS-SEC variable is created independently by another data manager – the questions are in line

with other surveys, so an independent check is carried out to ensure that the syntax was correctly created. The checker also runs the syntax themselves to check that they can replicate the results in the data.

Checks on tables - Waves 1 to 6

The data tables for Food and You 2 were produced by Ipsos for Waves 1 to 6 of the survey.

In Waves 1 and 2, Ipsos produced the tables using Quantum and subsequent checks were run against the table specification, ensuring all questions were included, that down-breaks included all categories from the question, that base sizes were correct (e.g. for filtered questions), base text was right, cross-breaks added up and were using the right categories, nets were summed using the correct codes, and that summary and recoded tables were included. Once the tables were signed off, the SPSS dataset was exported from Quantum. Weighting of the tables was also checked by applying the correct weight on the SPSS file then running descriptives and cross-break tabulations to check that this matched up with the values on the tables. If any errors were spotted in the tables, these were then specified to the data processing team in a change request form. The data processing team then amended the tables based on this and the tables were rechecked after the changes were made. The data and table checks were carried out by a team at Ipsos.

In Waves 3 to 6, a new process was introduced which switched around the order of data production, with the SPSS data created and signed off first, and then production of the tables (using Quantum) done subsequently. All checks of the tables remained the same.

Checks on tables - Wave 7

From Wave 7, the data tables are produced by the FSA Statistics team.

Before the tables were produced, the table specification was checked for inconsistencies including duplicate rows or tables, bases or SPSS variables appropriate to the table definition.

Once the tables are produced, automated testing checked that:

- all required tables were present, and contained the correct response columns, net columns and demographic breakdowns
- where the same base was used for different tables, results in the Total Base column were consistent between them
- for questions where respondents pick a single option, the percentages in a row totalled 100%
- for questions where respondents pick a single option and where the table had a net column, the relevant percentages summed to the values in the net column
- the Country breakdown in the all-country tables produced the same percentages as the Total rows for the individual country tables
- for each available breakdown, the unweighted totals in the three individual country tables summed to the unweighted total in the all-country table

The results for selected tables were reproduced separately and checked against those in the published tables. These tables were chosen to represent a variety of table types.