FY2 Wave 6 Technical report: Data validation and management

Overview

Questionnaire versions

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

In addition, the online data were available immediately in their raw form, however the postal questionnaire data must be scanned and keyed as part of a separate process. Tick box answers were captured by scanning, and numbers and other verbatim answers were 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 was taken as the basis for data processing. Once that was processed then a data map/dictionary was used to match the data from the postal questionnaires with the online data.

A wide range of edits were 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 were subject to errors introduced by participants and subsequently edits were required for this data. There are five key principles to editing postal data which were drawn upon for this:

1. Forward editing was applied to all filtered questions. If a participant was eligible to answer a question but had not, they were assigned a code of -99 “Not stated”.
2. A small number of back edits were applied to a handful of variables. If a participant had answered a question but had not answered “yes” at the previous filter question a back edit was applied. This was only done on variables specified by the FSA as the forward editing approach handles the majority of the cleaning required.
3. A specification was created by the FSA that set out a number of variables which needed to be edited to directly match the online routing. This was applied as a post field edit to the postal data only.
4. If a question was incorrectly answered as a multi-code question then the responses were set to -99 “Not stated”.
5. On a handful of questions that offered a multi-code answer but we asked participants to limit their answers to a maximum of three – answers were randomly assigned by running a random selection in SPSS. This was run for participants who answered more than 3 answers and the process ensured no duplicate answer could be selected.

In addition to this, where there was a multi-code variable that also had an exclusive code (such as “don’t know”), answers were edited so that valid multi-code options took priority, and conflicting exclusive codes were deleted. Where there were several exclusive codes, a hierarchy was 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 and 6, 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

Some cases were removed from the data if the participant completed both the online and the postal survey. In these instances, the online questionnaires were prioritised as that represents a more complete set of data. A total of 95 duplicates were removed from the data.

Coding

Coding was 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 were able to complete the survey in English and in Welsh. There were a small number of participants who chose to complete the survey in Welsh and provided verbatim text. These verbatims were 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 was then updated throughout the analysis process of Waves 2-6 to ensure that any newly emerging themes were captured. Developing the coding framework in this way ensured that it would provide an accurate representation of what participants said. This process was continued in Wave 6, with the codeframe developed further to match newly-emerged themes. After adding in any new codes to the codeframe, it was then reviewed by the FSA and Ipsos research teams with queries subsequently addressed by the coding team. After this it was then appended to the datasets.

Codes were grouped together into broad themes (for example, ‘Environmental and Ethical Concerns’), shown in bold text in the data tables. Some of the broad themes also had sub-themes (for example, ‘Fair Trade / Ethical’). For consistency between waves, all codes developed for the Waves 1-5 codeframes were included in the Wave 6 codeframe, including codes for which no responses were assigned at Wave 6. These codes are also present in the Wave 6 tables (and are marked as having received no responses).
Ipsos used 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 consultation projects. Responses were uploaded into the Ascribe system, where members of the Ipsos coding team then worked systematically through the comments and applied a code to each relevant piece of text.

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

Data checks

Checks on data

Ipsos checked 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 was checked a list is produced that identifies which variables require an edit and this largely related to the postal data. Any edits applied are set out in the section on 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 were created in syntax and are based on the table specification. All derived variables were checked against previous waves to ensure the values were roughly in line with what we would expect to see. Cross checks were carried out on the syntax used to create the derivations to ensure the logic was valid.

Once the derivations were set up the dataset was checked by other members of the team. Some derived variables were based on one question (for instance age) and these were checked by running tabulations on SPSS from the question they were derived, to check that the codes fed into the groups on the cross-breaks. If the derived variables were more complex and based on more than one question, for example, NS-SEC, more thorough checks were carried out. For example, the NS-SEC variable was created independently by another data manager – the questions are in line with other surveys, so an independent check was carried out to ensure that the syntax was correctly created. The checker also ran the syntax themselves to check that they could replicate the results in the data.

Checks on tables

Once the data was signed off the tables were produced using Quantum and subsequent checks were run against the table specification. These checks ensured all questions were included, that

down-breaks included all categories from the question, that base sizes were correct (for example, 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. 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 of five people at Ipsos.