WP.18
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

Editing Challenges for New Data Collection Methods

Prepared by Rachel Skentelbery & Carys Davies, Office for National Statistics, United Kingdom

I. Introduction

1. The Office for National Statistics (ONS) is currently investigating ways to improve data collection, with the main aims of making it easier for respondents to submit data to us, to reduce costs and make quality gains. There are a number of options going forward including Online Electronic Data Collection (eQuestionnaires), making better use of administrative data accessed from other organisations, and obtaining data fed directly from businesses (direct data feeds). There are also a number of challenges with editing for all of these different collection modes.

2. Currently the majority of our business data are collected through one of three modes; paper based surveys, telephone data entry (TDE) and offline collection instruments i.e. typically spreadsheets downloaded and therefore completed offline, sent via secure electronic file transfer (SEFT) systems. TDE is mainly used for surveys with a small number of questions, for example our Monthly Business Survey (MBS). Respondents can dial a number and enter their data over the telephone. This mode has proved quite popular and answers are automatically read back to respondents to ensure they have given the correct data. This can reduce some of the edit failures that occur, however, at present the majority of the editing is still run back at ONS so respondents may still be called back to amend or confirm their data. Offline collection instruments are mainly used for large companies with a lot of data to return, and more generally for complicated surveys, for example Foreign Direct Investment (FDI). These instruments will contain some basic edit rules which may correspond to formatting within the spreadsheet, or check components add to totals. This initial set of edits makes it easier for the respondents to complete and also means they are less likely to be called back. However, another set of edits are run back in the ONS and teams may still clarify variable values with the business via the telephone. Some large companies have set up programmes within their systems to automatically transfer their data to the collection instrument, which can help to alleviate respondent burden. Finally, paper based surveys are used for the majority of our business data surveys. Responses have to be edited at ONS and we utilise automatic editing (for £000 errors and totalling), selective editing and traditional edit rules. This can place a burden on our respondents who will be re-contacted, sometimes more than once, if their data appears to be suspicious or in error.

3. The majority of our social surveys are collected through Computer Assisted Personal Interviewing (CAPI) or Computer Assisted Telephone Interviewing (CATI). Within these modes edits are built into the questionnaire that the interviewer accesses. Any erroneous or suspicious values are flagged to the interviewer who can ask the respondent for clarification. They can then either correct the response or add a comment. The completed questionnaires are then securely transferred back to ONS.

4. ONS is seeking to invest in an Electronic Data Collection Programme (EDC), which aims to develop systems, methods and processes to improve the collection, integration and analysis of data. A phased approach is proposed; the first phase will put the core architecture and functionality in place. It will also develop secure electronic messaging and further develop offline collection instruments. Subsequent phases will aim to develop the methods, and enhance the systems and interfaces required for delivering more complex eQuestionnaires for both business and household surveys.

5. There are further potential opportunities to enhance the platform to facilitate the transfer of administrative data, in particular direct data feeds from businesses' accounting, payroll and human resource systems.

6. This paper will highlight the editing challenges in the different data collection modes and will focus on how to deal with edits in eQuestionnaires. It will discuss the experimental design we will be using to test edits in e-questionnaires in terms of presentation and the extent to which they should be utilised.

II. Data Collection modes

A. Administrative Data from other government departments

7. ONS has a series of work packages to investigate the use of administrative data for our economic estimates. One of the strands of this work is to investigate how to validate and clean the administrative data which have recently been accessed via data sharing order. These datasets include VAT Turnover and Expenditure (Purchases) data. ONS are also currently seeking access to data relating to Company Accounts, Corporation Tax, Income Tax and National Insurance Contributions. There are two main challenges when it comes to editing administrative data:

a) The administrative datasets are very large. This can make it very difficult to make manual changes because the number of suspicious values can be too large in number for the available resource. A consequence could be that it would be less efficient to do this than running the paper based surveys we currently have. Therefore it will be important that the majority of changes made to the data be carried out automatically.

b) In the UK ONS cannot telephone businesses that have suspicious values within administrative data. This is because the data has been collected for a non-statistical purpose. We can have some influence on the agency collecting the data by working with them to ensure that the administrative source is of sufficient quality; however we will need to clean the data once it is received at ONS.

8. Lewis & Lewis (2011) propose a number of methods for detecting and correcting errors and suspicious values in VAT Turnover data. These methods focus on looking for suspiciously small or large values in the dataset, as well as suspicious patterns and unit errors. Lewis & Finselbach (2012) utilise these methods within the context of the Monthly Business Survey to understand whether the cleaning of administrative data needs to be specific for its end purpose. We are also currently investigating whether these methods are appropriate for VAT Expenditure data and, at a later date, Company Accounts data.

B. Administrative Data sourced directly from business systems (Direct Data Feeds)

9. Direct Data Feeds are when a company can submit its data directly to ONS from payroll, accounting and Human Resources (HR) systems. Other countries have already implemented Direct Data Feeds with some benefits e.g. CSO Ireland with respect to payroll data. However, for ONS there will need to be some work before this collection mode becomes a reality. At present, we have not investigated the editing challenges for this approach as no data is available. However, most of the challenges will be the same as with administrative data, as this source is really administrative data accessed via an alternative, more direct channel i.e. the datasets will be extremely large which could lead to issues with manual changes, however we can assume that as the data is obtained directly from companies we should

be able to recontact them if required. The expectation is that this data should be relatively precise as it will be coming directly from the companies systems, however without further investigation we cannot be sure. This work will be taken forward shortly, once funding is available.

C. Online eQuestionnaires

10. To date ONS has used Internet data collection methods for two of our business surveys, Occupational Pension Scheme Survey (OPSS) and a pilot for Capital Expenditure (Capex). On the social survey side the Opinions survey (OPN) and Labour Force Survey (LFS) have also been piloted. So far the edits within these collections instruments have been limited and any gains have mainly been due to automatic routing within the questionnaires. No query (soft) edits have been tested in the questionnaire so far, only fatal (hard) edits have been included. The next phase of the EDC programme will include a work package to research the use of edits within online questionnaires in order to create initial standards and guidance for online editing for both business and social surveys.

11. There are a number of challenges in online questionnaire editing and a set of questions that we would like to investigate:

a) What type of edits should be used in an online questionnaire? Should just fatal edits be used or are query edits also useful?
b) How many edits are too many? Will lots of edit failures discourage response?
c) Should it be compulsory for all edits to be resolved, or commented upon, before submission of the data is possible?
d) How should edits be presented to respondents for maximum effect?

12. Initially we have carried out a literature review to ascertain what other countries have done, as well as looking at web design standards. Results of this are summarised in section III.

III. Summary of Literature for Online Edits

13. *What types of edits should be used in an online questionnaire?* - In the Canadian 2011 Census mainly consistency checks, checks for invalid responses and checks for partial and non-response were used (Laroche, 2011). Some countries, for example Spain and US Census Bureau, went a little further and added range edits to the questionnaires. In order for range edits to be used the edits first need to be tested using previous data and discussions with respondents, to ensure that the range is valid. There were some cases in the US Census Bureau's usability testing where respondents set their answer to the upper bound of the range in order for it to pass validation (US Census Bureau, 2005). This is something that must be avoided. Some countries have more types of edits than others, however in order to save money the feeling is that as much validation as possible be moved to the questionnaire. This, however, should only occur with research beforehand to ascertain how respondents react to various edit rules.

14. *Can you have too many edits in an online questionnaire?* – One concern is that if respondents have to trawl through lots and lots of edit failures they will either stop filling in the questionnaire and become non respondents, or falsify their answers in order to pass the checks. Statistics Canada point out that this was a concern in their 2011 Census (Laroche, 2011) however the US Census Bureau (2005) make a very good point that it doesn't matter how many edits you have in the questionnaire, it is the likelihood of a respondent triggering that check that it important. Their experience generally shows that respondents are receptive to edit checks; people are getting used to filling in forms online and expect to be prompted if something is wrong.

15. *Should it be compulsory for all edits to be resolved, or commented upon, before submission of the data is possible?* – There is a risk that if it is compulsory to resolve edits before submission then respondents will not submit and unit response could increase. The US Census Bureau (2005) point out that it is better to obtain some data, even if they fail edits than no data at all. The Energy Information Administration's (EIA) Internet Data Collection System included both query and fatal edits. In order to

resolve the fatal (hard) edits, and submit the survey, respondents either had to correct the appropriate item or provide a comment on the data. For query edits, respondents could correct, comment or take no action (Barlow et al, 2008). Other countries have 'mission critical' variables which respondents have to provide in order to be able to submit the questionnaire. In the UK 2011 Census the list of people within the household had to be completed by respondents, this was because the rest of the online questionnaire was generated as a result of this response. Therefore it was critical that a response was obtained.

16. *How should edits be presented to respondents for maximum effect?* – Different National Statistics Institutes (NSIs) have differing approaches and views on this. There are a few options for the timing of edits; immediately after an item is entered, after a section of questions are completed and after the whole survey has been completed. The US Census Bureau recommends that edits should be performed immediately as study participants preferred this. Immediate edits allow the respondent to correct the error straight away and can help to avoid similar mistakes later on. However, this approach can cause issues for edits which include more than one variable e.g. components adding up to a total. Therefore there is a necessity for these edits to be run once all relevant questions have been completed. The 2006 Canadian Census displayed edits screen by screen, each of which displayed a group of related questions (Laroche, 2008).

17. Another option is to provide a list of error messages once the survey has been completed. This could cause some issues with navigating back through the questionnaire to make corrections. The 2009 Spanish Agriculture Survey implements edits in two stages. Errors are detected automatically while the respondent is filling in the questionnaire. Respondents can either resolve the errors immediately or ignore and continue. A list of uncorrected errors is presented at the end of the survey, giving respondents an opportunity to make corrections (Revilla et al, 2009).

18. In addition to the timing of edit check flags within the questionnaire it is important to consider the appearance of the flags e.g. pop up boxes, text beside the question relating to the edit or a list of failures. The positioning of error messages is dependent on the timing of the checks but also on the layout of the questionnaire. There are two common online survey designs; paging and scrolling. Paging surveys are where questions appear on the screen page by page. Scrolling surveys are where the survey is seen as one long scrolling page. Couper (2008) suggests that paging surveys are far less likely to suffer from omission errors than scrolling surveys.

19. When using a scrollable design the US Census Bureau found that presenting errors at the top of the page means that respondents have to scroll up to see a list of errors, which was generally disliked. The use of pop up boxes to display error messages, with a scrollable design, can sometimes mean that the item which triggered the edit could be off the viewable screen (US Census Bureau, 2005).

20. Most of the literature reports that the error messages used in online surveys are displayed in red text in order to make them stand out from other text. Some NSIs also use larger text than the question wording.

21. Schonlau et al (2002) suggest that it is ideal for error messages to be placed directly above or below the item which requires attention. The message should at least note where the error has occurred and the nature of the problem. The wording of the message should be polite, specific and should place no blame of the respondent. The use of generic error messages are not helpful as they can lead to confusion as to what action needs to be taken in order to resolve the error (Couper, 2008). Participants in a study by the US Census Bureau (2005) commented that error messages are more useful if they detail the item that needs attention and the specific action, we should not assume that a respondent will know what to do when an error is flagged. Mockovak (2005) also completed a study which presented respondents with three different designs for the edits. The results showed that users expressed a clear preference for the error message to appear under the item; however timing was not as important to them.

IV. Testing Proposals

22. In order to answer some of the questions raised in the literature review we will be carrying out some testing in winter 2012. The initial study will be run alongside our Data Collection Methodology Centre of Expertise (DCM) and will utilise cognitive interviewing. The approach will test edits alongside different questionnaire designs. DCM have decided that only paging designs will be tested which is a good first step to ensuring the edits work well. Their research has also shown that paging surveys appear to be the best option for obtaining quality data. The project aims to:

- Ensure edit checks are understood by respondents and clearly describe the error and what to do to correct/ suppress; reducing the risk of respondent error.
- Consider number of edits to ensure that non response isn't introduced.
- Consider inclusion of query and fatal edit checks.
- Consider how edits should be presented to respondents.

23. The initial study will concentrate on the number and types of edits that will be included in the questionnaire. The pilot will be tested on a household survey, the Labour Force Survey (LFS). Two edit set ups will be tested:

Edit Set up 1:

1. Error messages appear alongside the variable it relates to and will appear in red text immediately after the respondent has completed the question.
2. If the edit relates to a group of questions then these questions will all appear on the same page and the error will be shown immediately after the last question is answered.
3. If an answer is not changed following an error message, respondents will have the opportunity to add a comment.
4. Any uncommented and unresolved errors will also be presented as a list before submission, with the option to ‘correct’.
5. All appropriate query and fatal edits from the current LFS will be included.

Edit Set up 2:

- As 1, 2, 3 and 4 above but with only fatal edits flagged to respondents.

24. In the initial test the only differences will be the number and types of edits used. Respondents will be observed to see how they react to the various edits i.e. are they noticed? do the respondents take action? If respondents do not take action we will question why this is the case. We will also seek respondent feedback on how the edit design could be improved. We will utilise DCM’s expertise in cognitive interviewing to ensure these questions are answered.

25. After this initial testing we will create an improved design, taking into account respondent feedback and what we have learned from the two set ups. We should be able to take a decision on whether query edits can be included and how many edits to include. From this a new test scenario can be proposed and further testing can take place.

V. Future Research

26. Once the initial testing for the Labour Force Survey is complete, an initial business survey pilot will also take place with the aim to developing a set of ‘Standards & Guidance’ for editing within web questionnaire by 2014. We will, of course, also take other countries research into account when putting these standards together and would welcome feedback from other NSIs on their experience with editing internet surveys. At the same time, work will also carry on checking and correcting administrative data. By 2014 we aim to also have a set of standards and guidance for how to create a ‘clean’ administrative dataset.

References

Barlow, B., Freedman, S., Weir, P. "**Data editing in a common internet data collection system".** Proceedings of UNECE Conference of European Statisticians, Vienna, Austria, April 2008. Web: http://www.unece.org/fileadmin/DAM/stats/documents/2008/04/sde/wp.3.e.pdf

Couper, M,P. (2008) **Designing effective web surveys.** Cambridge University Press.

Laroche, D., Grondin, C. **"Impact of online edits and internet features in the 2006 Canadian Census".** Proceedings of UNECE Conference of European Statisticians, Vienna, Austria, April 2008. Web: http://www.unece.org/fileadmin/DAM/stats/documents/2008/04/sde/wp.5.e.pdf

Laroche, D. "**The evolution of edits in the Canadian Census of population online questionnaires".** Proceedings of UNECE Conference of European Statisticians, Ljubljana, Slovenia, May 2011. Web: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp.10.e.pdf

Lewis, D. and Finselbach, H. **"Editing and Imputing VAT Data for the Purpose of Producing Mixed-Source Turnover Estimates".** Proceedings of UNECE Conference of European Statisticians, Oslo, Norway, September 2012.

Lewis, D. and Lewis, P. "**Finding and Resolving Data Quality Issues in Value Added Tax Data"**, final report for WP2 of ESSNet on Admin Data, 2011.

Mockovak, B. **"An evaluation of different design options for presenting edit messages in web forms".** Bureau of Labour Statistics, 2005.Web: http://www.fcsm.gov/05papers/Redline_IXB.pdf

Revilla, P., Arbués, I., Saldaña, S. **"Editing multimode data collections: The Spanish experience".** Proceedings of UNECE Conference of European Statisticians, Neuchâtel, Switzerland, October 2009. Web: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2009/wp.12.e.pdf

Schonlau, M., Fricker, R. D., Elliott, M. N. **"Guidelines for designing and implementing internet surveys (chapter five)"**. Conducting research surveys via email and the web. Web: http://www.rand.org/content/dam/rand/pubs/monograph_reports/MR1480/MR1480.ch5.pdf

U.S. Census Bureau, **"Designing interactive edits for U.S. electronic economic surveys and censuses: Issues and guidelines".** Proceedings of UNECE Conference of European Statisticians, Ottawa, Canada, May 2005. Web:http://www.unece.org/fileadmin/DAM/stats/documents/2005/05/sde/wp.22.e.pdf