

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (ii): Editing nearer the source

EDITING IN MIXED MODES IN THE 2011 ENGLAND AND WALES CENSUS

Supporting Paper

Submitted by the Office for National Statistics, UK¹

I. INTRODUCTION

1. For the first time in England and Wales, the 2011 Census of Population will offer households the option of completing an internet questionnaire as an alternative to paper self-completion. It is expected that about 25 per cent of the population will choose to complete the 2011 Census on-line.

2. Offering the Census questionnaire on-line represents a new experience for editing at ONS. Since paper self-completion represents the dominant mode of collection, the internet questionnaire has been specifically designed to follow the paper version as closely as possible. However, at the same time, ONS has sought to utilise and exploit a range of computer and web attributes in order to maximise data quality and minimise respondent burden. Thus, there is a need to fully understand what the implications are for the editing process of using the proposed computer and web attributes.

3. The 2009 Census Rehearsal takes place in October 2009 and is a major test of the systems and operations to be used in the Census proper. Whilst a small amount of usability, accessibility and cognitive testing have already been completed, the Rehearsal provides a unique opportunity to test the internet environment in a fully operational setting.

4. This paper follows on from Wagstaff and Wallis (2008) which discussed first thoughts on editing in mixed modes in the 2011 Census. Section II of this paper provides a high level overview of the 2011 UK Census Editing Strategy and the broad principles which were followed when developing the internet questionnaire. Section III describes the 2009 Census Rehearsal including the rationale for the selected areas and a description of the high level design. Section IV walks through the web technologies being applied in the Rehearsal and discusses the implications of each for the editing process. Finally, Section V provides some brief concluding remarks.

¹ Prepared by Heather Wagstaff and Sue Dalton (Heather.Wagstaff@ons.gov.uk Sue.Dalton@ons.gov.uk)

II. 2011 UK CENSUS EDITING STRATEGY

5. The 2011 UK Census Editing Strategy has been developed with the primary aim of imputing for all missing data and resolving inconsistencies in the responses for the households and persons affected. The Strategy is driven by three key principles:

1. all changes that are made will maintain the quality of the data;
2. the number of changes to inconsistent data will be kept to a minimum; and
3. as far as possible, missing data should be imputed for all variables, to provide a complete and consistent database.

6. In addition, a number of broad principles were followed when developing the internet questionnaire in order to ensure a high quality design:

1. no unnecessary changes will be made from the paper questionnaire;
2. respondent burden will be minimised by making the user experience for internet data collection (IDC) as simple and easy as possible.
3. adherence to web design best practice where possible;
4. ensuring accessibility for everyone including those who require assistive technologies when using the internet.

III. 2009 CENSUS REHEARSAL

7. The 2009 Census Rehearsal will be held on 11th October 2009. The key objective is to pilot the field operations and processing activities for the Census proper. The Rehearsal has been designed solely for the purposes of operational readiness and provides a unique opportunity to test the internet environment in a fully operational setting. Any changes to the 2011 Census design that are not piloted during the Rehearsal will undergo separate small scale testing.

8. In order to identify the Rehearsal areas, a series of criteria were developed and applied to each Administration Area in England and Wales. These criteria ensured that the Rehearsal would be representative of the population and that the selected Areas would not be overburdened with field testing. The procedure resulted in the selection of four Administration Areas comprising about 135,000 dwellings which, together, were considered demanding enough to pilot the enumeration plans. The Rehearsal Areas and the rationale for their selection are as follows:

Lancaster is a large town which contains an above average number of flats, a prison, a small military base, several hospitals and student halls of residence. There is above average internal population movement combined with an above average population growth. Lancaster represents a typical workload for a Census Area Manager and enables the rehearsal of recruitment and local management arrangements.

Newham is a Borough Authority within inner city London and is part of the Greater London Authority. The area represents the harder to enumerate population and includes a large and mixed ethnic minority population with an above average inward migration of people from the European Union (EU). There are an above average number of flats and multi-occupied addresses. The area allows refinement of assistance, translation and coverage procedures.

Isle of Anglesey is in coastal rural north Wales with a very high proportion of Welsh speakers. It has high number of students from a nearby University, an Air Force base and high numbers of holiday/second homes. Anglesey was identified as a good location for piloting recruitment and

localised publicity. It allows refinement of Welsh language translation procedures and the Census Contact Centre.

Birmingham is a large inner City with a mix of ethnic diversity. It was included for the expected complexity of the field operations, but for reasons of timeliness and cost, the data will not be captured and coded.

9. The population will be counted at their place of usual residence in England and Wales regardless of their actual location on Census Day. ONS is currently developing an address register which will identify and manage the enumeration of dwellings. Each physical dwelling will be assigned a unique identifier and unique Internet access code. Each paper questionnaire will be pre-printed with the dwelling address, their unique identifier, their Internet access code and the Census website address.

10. The enumeration method is principally mail-out/mail-back: questionnaires will be mailed to the majority of dwellings in the Rehearsal Areas; questionnaires for a small proportion of households, which are deemed difficult to contact, will be hand delivered by enumerators. The Rehearsal will trial eleven different questionnaires in two languages (English and Welsh).

Testing the Online Questionnaire

11. The Rehearsal will also provide an opportunity to further test the internet questionnaire in a fully operational setting. During its development, the online questionnaire has been subjected to three waves of usability and accessibility testing and also cognitive interview based testing. The Usability and Accessibility testing was conducted by Web Technologies Group (WTG) a sub-contractor within the Lockheed Martin 2011 Census Consortium, with input from AbilityNet UK when testing with disabled users. AbilityNet UK is national charity that helps disabled adults and children use computers and the internet by adapting and adjusting their technology. Cognitive interview testing was conducted by ONS Data Collection Methodology (DCM).

12. The aim of the usability testing was to identify problems experienced by respondents using a prototype of the Census online questionnaire. The focus of this testing was on the respondents ability to successfully use the system for data collection. Testing was performed with respondents who had a range of different characteristics. Respondents were asked to complete parts or all of the questionnaire using pre-specified scenarios rather than as themselves. Following analysis of the findings, recommendations were put forward for implementation and further testing at subsequent waves.

13. The Accessibility testing was carried out by AbilityNet UK and focused on scenario-based tasks performed by disabled user groups who use a range of techniques and assistive software. Assistive software and techniques tested included: voice recognition; screen readers for blind people; screen magnification for those with poor eyesight; key board only users; and colour contrasts for dyslexic users. The issues identified related to accessibility of the system when using assistive software and recommendations for change were implemented to ensure the system complied with guidelines that explain how to make web content accessible to people with disabilities (W3C (1999)).

14. Whereas the Usability and Accessibility testing focussed on interaction with the system, the cognitive interview based testing conducted by DCM aimed to identify potential mode effects of the Internet questionnaire. A short prototype containing single examples of the various question style types (radio buttons; check boxes; text entry; and drop-down boxes) was used to identify potential sources of measurement error due to mode effects. Some errors arose with the use of drop-down boxes in the relationship matrix which were not identified in either the

usability or accessibility testing. There were also some problems identified with the qualifications and the ethnicity questions where the complete set of responses were not visible without scrolling down the page. The ethnicity question has now been updated to include a line to remind respondents that more options are available. Other issues arose with people missing instructions.

15. The 2009 Rehearsal will further test the system. A stakeholder review, a short evaluation questionnaire for respondents and further DCM cognitive interview based testing are being planned to be conducted during the nine week period whilst the 2009 Rehearsal questionnaire is live.

IV. INTERNET CAPTURE IN THE 2009 REHEARSAL

16. The introduction of IDC for the 2011 England and Wales Census provides an opportunity to improve data quality and reduce respondent burden. The design of the Internet questionnaire has followed the key principle that no changes will be made from the paper questionnaire unless they improve the user experience and data quality whilst simultaneously seeking to minimise mode effects. This is inline with the recommendation that a unimode design will reduce the risk of mode effects (Dillman (2007)). Internet standards will be followed; for example, check boxes will be used to indicate that multiple responses are possible and radio buttons where only one response is possible.

17. A literature review of international experience from the 2005/06 Census round indicates that Internet users are highly computer literate and expect to see validation checks such as radio buttons and check boxes. Evidence suggests that it would seem to be implausible to have no validation. In fact, the rationale for validation checks and automated skipping is integral to good questionnaire design (Laroche (2005), Statistics Canada (2007), Laroche and Grondin (2008)).

18. The ONS carefully considered the potential implications of each change before introducing anything in the Internet questionnaire that was different from the paper version. For the Rehearsal, ONS have achieved a balance between general adherence to the paper version while incorporating a number of useful web standards. In the remainder of this section we consider the implications for editing which arise as a consequence of the Internet questionnaire as it will be used in the 2009 Rehearsal. We discuss the various question style types being trialled with reference to potential sources of measurement error due to mode effects. Firstly we consider personalisation techniques, followed by on-line validation, automated routing, radio and check buttons. We close by describing some of the minor changes to question wording.

Personalisation Techniques

19. When completing the Rehearsal questionnaire on paper, the respondent is asked to record the names of all usual residents three times: firstly, in the listing grid; secondly, in the relationship matrix; and thirdly, at the start of each set of person questions. Analysis of a sample of 2001 responses identified instances where the ordering of names was not consistent across the three listings which resulted in one persons responses being recorded in the record space for another. This was especially marked amongst the larger households. Further, a number of relationships were recorded the wrong way round, for example parent instead of child. The application of personalisation techniques in the on-line interface goes some way towards possible mitigation of such errors.

20. When completing the Internet questionnaire, the respondent will be asked to record the number of usual residents and write their names in the listing grid. This information is used by the electronic interface to firstly, create the appropriate number of sets of individual questions and then to populate the name fields in the relationship matrix and on the individual questions.

However, there will be modal variation amongst the collection instruments which is explained by considering the relationship matrix.

21. As in 2001, the 2009 relationship matrix seeks to identify family structure and hidden families by capturing the relationships between household members. Briefly the form-filler is asked to record the name of each household member and then indicate the relationship of that person to other persons in the household. The question occupies a double-page spread and is headed by a simple worked example. The question is formulated, “How is Person 2 related to Person 1” as shown in Figure 1.

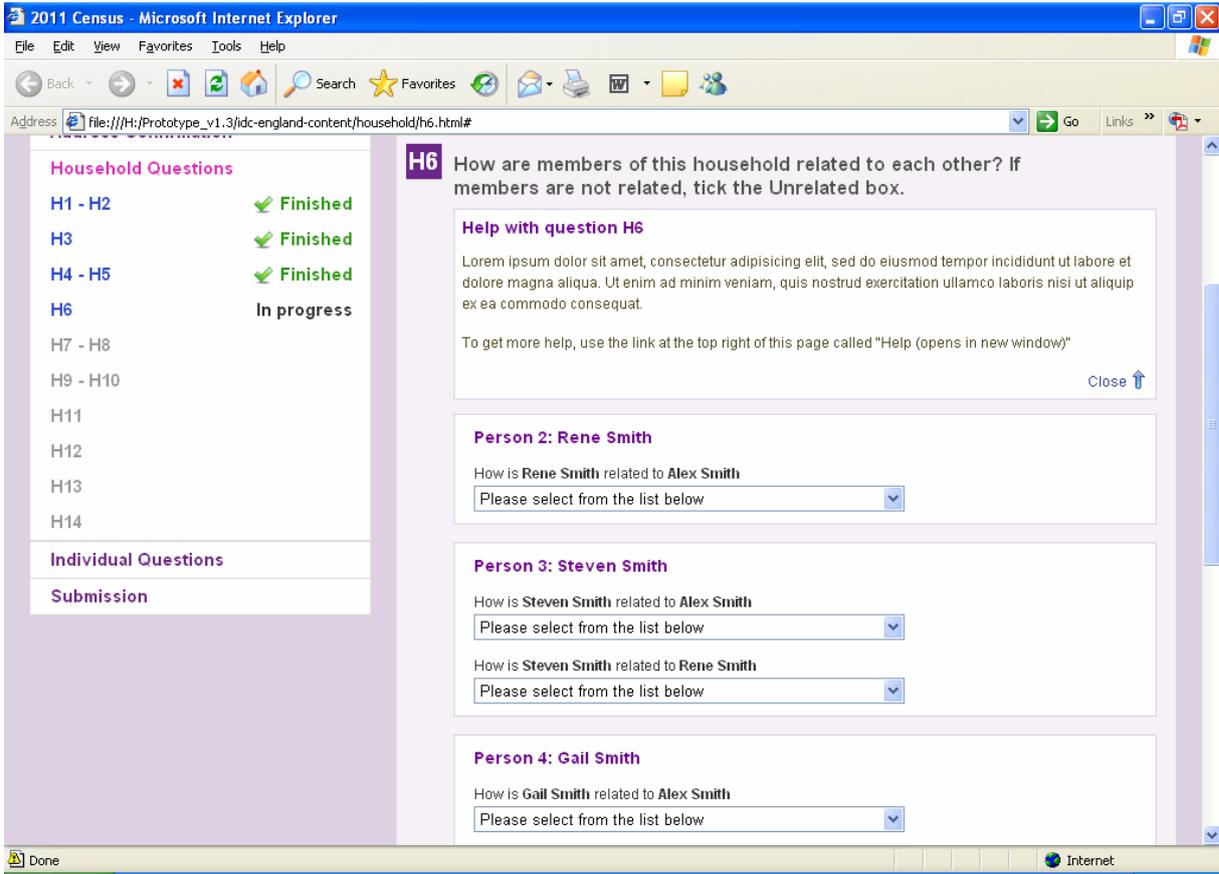
Figure 1: Relationship Matrix – Paper Collection

Source: 2009 England and Wales Census Rehearsal

22. As explained earlier, the format of the on-line relationship matrix has been amended to make it easier to complete and to improve data quality by personalising the question for each respondent as shown in Figure 2. As an aside, Figure 2 also clearly shows the navigation bar at the left hand side of the screen which informs the respondent how far they have progressed through the questionnaire and allows navigation between questions and from one person to another.

23. The on-line interface will significantly improve data quality by writing the names recorded on the listing grid directly to the relationship matrix. The question will appear in the form, ‘How is Rene Smith related to Alex Smith’ above a drop-down menu which displays the full set of response options. Conversely, the on-line questionnaire does not automatically display a worked example to aid the form-filler who must actively seek help via the on-line help facility.

Figure 2: Relationship Matrix – Internet Collection
 Source: 2009 England and Wales Census Rehearsal



24. The relationship matrix is the most complex question to complete and as household size increases so the number of relationships increase. The use of personalisation and drop down menus are techniques directed towards improving data quality and may mitigate against the ordering and reporting errors identified in 2001. However, there are clear differences in the provision of help between modes, the effect of which must be carefully assessed.

25. Large households represent a small population which are important to identify and enumerate for purposes of resource allocation. In 2001, less than 2% of all households returned a Household/Continuation Form combination. There is concern that large households who complete the paper instrument may be under enumerated since households of size seven and over must actively request one or more Continuation Forms. The electronic interface automatically creates records for up to 30 persons, and therefore large households will be encouraged to use the Internet. However, evidence from 2001 shows a clear break in data quality at the threshold of the Continuation Form which is especially marked in the quality of relationship information. If large households are related to poor literacy, for example, then it may be that respondents become confused and frustrated by the volume and complexity of the on-line relationship questions, break-off early in the process and not realise the requirement to complete the paper questionnaire.

On-line Validation

26. The on-line interface will apply a sub-set of preliminary edits in real time. However, the primary concern associated with the implementation of on-line validation surrounds issues of public perception. The experienced internet respondent expects to see some degree of validation and might question the credibility of the questionnaire if validation was not present. Conversely,

the inexperienced internet respondent might be unable to work their way through even simple validation and we might observe an unexpectedly high drop-off rate from such respondents.

27. For 2009, validation rules will be applied throughout the Internet questionnaire in an attempt to maximise data quality. Rules in respect of missing, implausible and inconsistent responses are mainly soft: if a respondent fails a soft rule then a warning message will be posted to the screen alerting them that they have not completed the response as requested. However, they can choose to ignore the warning and continue through the questionnaire without correcting the error. However, cognitive testing identified the fact that some respondents, when viewing validation messages, thought that they had to provide an answer in respect of soft validation in order to continue through the questionnaire. International experience suggests that in general, validation messages are effective in obtaining answers to questions that respondents might otherwise have overlooked and in having respondents correct errors they inadvertently committed (Laroche and Grondin (2008)).

28. In terms of internet capture, hard validation stops the respondent progressing through the questionnaire until they satisfy the rules for that question. In 2009, hard validation will only be applied to four key questions which are essential for the electronic interface to construct the questionnaire. The questions relate to the number of usual residents and visitors which will allow the interface to construct the correct number of sets of questions to be completed for all residents and visitors. Therefore, it is critical that these questions are completed correctly. If they are not then the respondent must revert to completing the paper questionnaire.

Automated Routing

29. Automated routing is being used to reduce respondent burden. Where automated routing is applied, questions which are not applicable are not displayed. For example respondents aged less than 16 years are not eligible to complete the Labour Market questions. Since the questions are numbered sequentially, a message is posted to the top of the screen for the next applicable question to indicate that one or more non-applicable questions were skipped. The progress bar is also updated with 'Not needed' for the skipped questions.

30. However, if respondents make an error in the routing question then they may never realise their mistake. Whereas on the paper questionnaire, because the skipped question is clearly visible, there is opportunity to identify the error. For example, if the respondent was aged 61 but erroneously typed 16 they be erroneously routed past the Labour Market questions. The respondent may not read the explanatory message associated with the routing and hence be unaware of their error. Analysis of 2001 found respondents sometimes did not follow routing instructions on the paper questionnaire, which resulted in increased respondent burden (completing questions not relevant to them) and additional processing load to identify and correct the error.

Radio Buttons

31. Standard functionality being trialled in the Rehearsal includes the use of radio buttons for single selections from multiple choice questions. However, even this simple technique is not without issue. For example, a respondent can select an erroneous response, then legitimately change their mind and select a different response. However, once a response has been selected, the question cannot then be left intentionally blank. Whereas, with the paper questionnaire, the respondent is instructed how to negate their original response and/or can leave the question intentionally blank.

Check Boxes

32. Following internet standards, check boxes will be used to indicate that multiple responses are possible and for more complex questions such as ethnicity which allow a multiple 'tick' and

text response. Thus, respondents will be permitted to select multiple categories and a text response in line with the paper form. Interestingly, with check boxes the interface offers the ability to check and uncheck text boxes. Thus, conversely to radio buttons, a respondent can make a selection, change their mind, and leave the item intentionally blank.

Changes to question wording

33. IDC has led to minor changes to instruction wording, for example ‘tick all that apply’ has become ‘select all that apply’. Similarly, IDC has made some instructions unnecessary, for example, the paper questionnaire instructs people who live alone to bypass the Relationship Matrix whereas the electronic interface automatically bypasses questions which are not applicable (but posts a message to the screen advising the user that it has done so). Conversely it has been necessary to add an amount of assistive text, for example to allow a change to the number of usual residents if respondents realise, when entering the names, that they have too many or too few.

Post Rehearsal Analysis

34. The internet questionnaire goes live on 18 September 2009 with delivery of the captured and coded data for both modes scheduled for January 2010. A thorough analysis will be undertaken of responses from the two modes. This will include a comparison of the overall distribution of responses by mode; followed by comparisons of the patterns of item non-response and invalid values. Of particular interest will be evidence of errors introduced specifically by the use of standard Internet functionality where we would look for distributional disturbances from: the use of radio buttons or check boxes; automated routing; on-line validation and changes to question wording. Other things to look out for include: heaping of dates for example 1st January; a comparison of drop off effects between modes; and evidence for where ‘any answer’ has been recorded just to get through the soft edits; a high proportion of countries with short names or beginning with A’s and B’s.

V. CONCLUDING REMARKS

35. The introduction of IDC for the 2011 England and Wales Census offers the opportunity to improve data quality and reduce respondent burden. The design of the 2009 Rehearsal internet questionnaire has followed the principle that no changes will be made from the paper version unless it will improve the user experience and quality of data collected, whilst minimising mode effects. Following the Rehearsal, the data will be fully analysed to ensure that all design decisions have been successful and that any issues identified are addressed in time for the 2011 Census.

VI. REFERENCES

Dillman, D (2007) *Mail and Internet Surveys: The Tailored design method. 2007 update With New Internet, Visual and Mixed-Mode Guide*. New Jersey, John Wiley and Sons Inc.

Laroche, D. (2005). United Nations Economic Commission For Europe – Statistical Data Editing, Volume No. 3, Impact on Data Quality, 2006, “Evaluation Report on the Internet Option of the 2004 Census Test: Characteristics of the Electronic Questionnaires, Non-Response Rates, Follow-up Rates and Qualitative Studies”, Ottawa, 2005.

Laroche, D. (2007). Advisory Committee on Statistical Methods, “2006 Census Internet Data Collection”, Meeting No. 44, April 30th and May 1st, 2007.

Laroche D and Grondin C. (2008) “Impact of Online Edits and Internet Features in the 2006 Canadian Census”. Working Paper No.8, UN/ECE Work Session on Statistical Data Editing, Vienna.

Statistics Canada (2007) “Census on the Net”, Invited Paper UN/ECE Work Session, Astana, Kazakhstan.

W3C (1999) “Web Content Accessibility Guidelines V1.0” <http://www.w3.org/TR/WAI-WEBCONTENT/>

Wagstaff H.F. and Wallis R. (2008) “First Thoughts on Editing in Mixed Modes in the 2011 England and Wales Census”. Working Paper No.4, UN/ECE Work Session on Statistical Data Editing, Vienna.