WP.4

ENGLISH
ONLY

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

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Vienna, Austria, 21-23 April 2008)

Topic (i): Editing of data acquired through electronic data collection

FIRST THOUGHTS ON EDITING IN MIXED MODES IN THE 2011 ENGLAND AND WALES CENSUS

Supporting Paper

Submitted by the Office for National Statistics, UK[1]

I. INTRODUCTION

1. The 2001 Census estimated the population of England and Wales to be over 52 million, an increase of 4.2 per cent since 1991. The Census painted a picture of society that reflected the wide variety of demographic changes that took place during the latter part of the 20th Century (ONS, 2005). In England and Wales (EW) the Census of Population and Housing is conducted every 10 years. The next one will be held in March 2011. For the first time, the Office for National Statistics (ONS) will offer all households in England and Wales the option to complete their questionnaire via a secure internet based application. ONS have specified the following five objectives for the internet questionnaire: to improve response rates, both overall and amongst the hard to count populations; to meet public and stakeholders expectations; to improve data quality; to achieve rapid availability of data; and to achieve cost savings.

2. Providing the option of internet questionnaires is one of the major challenges that ONS is currently facing. Although both the paper and online questionnaires are alternate modes of self-completion, an electronic questionnaire might not truly replicate the experience of a respondent completing the traditional paper version. Data editing is traditionally a post-collection process and one of the most resource intensive aspects of any census processing operation. To date, editing in the EW Census has been dedicated to methodologies, approaches and algorithms which have focused on defining edit rules and building editing systems with the goal of identifying and correcting for respondent error. However, internet capture offers the possibility of interactive editing to improve the quality of the responses, a feature which does not exist for paper self-completion. ONS is carefully considering how far on-line editing, including simple and complex coding, should be built into the electronic interface. It is difficult to predict or measure the impact that online capture might have on data quality given the self-selective nature of the population that might complete the Census questionnaire online.

[1] Prepared by Heather Wagstaff & Ruth Wallis (Heather.Wagstaff@ons.gov.uk, Ruth.Wallis@ons.gov.uk)

3. This paper discusses first thoughts on data editing in mixed modes in the 2011 Census. Section II sets the scene by briefly describing the editing process applied in the 2001 Census. Here we use 2001 as a proxy for 2011 since the questionnaire and editing processes are not fully specified. Section III discusses additional challenges for internet questionnaire design whilst Section. IV does similarly for data capture and coding. Section V discusses the integration of data from the two processing streams, whilst Section VI closes with some concluding remarks.

II. OVERVIEW OF THE 2001 CENSUS EDITING PROCESS

4 The 2001 Census Editing and Imputation Strategy was 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 followed four basic principles:

(1) all changes that were made would improve the quality of the data;

(2) the number of changes to inconsistent data would be kept to a minimum;

(3) as far as possible, missing data would be imputed for all variables to provide a complete and consistent database;

(4) the system had to be relatively easy to develop and capable of processing large amounts of data automatically within short timescales.

The editing process was driven by the format and content of the 2001 paper questionnaire and implemented at a number of points throughout the processing operation.

2001 Questionnaire

5. The style of the 2001 Census questionnaire moved away from the traditional matrix design towards a set of individually sequenced questions asking information about each person in the household. The latter style of questionnaire became known as the 'pages per person' format. The structure of the questionnaire was complex and contained three discrete sections: 9 questions about the household; a relationship matrix seeking information about family structure and hidden families; and 34 questions relating to each person. The person questions sought a mixture of quantitative and qualitative responses which were supported by six routing questions. The routing questions were developed to guide respondents through the questionnaire and to only complete those sections relevant to them. There were three types of routing question: simple routing which required no response; routing which required a single ticked response; and complex questions where the routing was combined with double barrelled questions which required both a tick and a textual response. One of the primary issues for editing was to ensure that individuals were correctly routed through the questionnaire based on their responses to key questions.

Data Capture and Coding

6. For the first time in England and Wales, the capture and coding of the 2001 Census data was outsourced to an external contractor under strict quality assurance criteria imposed by ONS. The completed questionnaires were transported to a single data processing centre situated in the North of England. The questionnaires were scanned and the data captured using Optical Mark Recognition (OMR) and Optical Character Recognition (OCR) software to translate the images into data records. The OCR software recognised both upper and lower case text and was specifically tuned to the EW style of handwriting. If characters were not recognised automatically, within a pre-determined degree of confidence, the image was presented to an operator for interpretation and keying. A set of univariate preliminary edits were developed to specify how the data should be captured and coded. After editing was complete, the Statistics Canada automated coding system, ACTR (Automated Coding using Text Recognition), was applied for complex coding where textual responses were converted into a numerical format. Responses which could not be coded automatically, or by computer assistance, were sent to expert coders to assign the appropriate code. An automatic quality assurance system was integrated within the capture and coding sub-systems to assess whether the coding was consistent and met pre-specified quality standards. For the whole of the UK, 18.5 billion tick boxes and 6.1 billion characters were captured from 27.3 million forms. Some 207 million tick boxes and 1.1 billion

characters were sent to keyers for correction, just over 1 per cent of the tick boxes and 17 per cent of characters captured.

Edit and Donor Imputation System (EDIS)

7. The fully captured and coded data was supplied to ONS for processing by a bespoke Felligi-Holt based Edit and Donor Imputation System (EDIS). The EDIS pre-processing sub-system applied multi-tick rules, range checks, constructed reciprocal relationships and ensured that responses were consistent with the question routing. The EDIS editing process scrutinised every household and person record to identify inconsistent fields. Inconsistencies were corrected within the edit phase where possible, or marked for subsequent imputation by a fully automated donor based imputation process. After imputation the consistency checks were re-applied to ensure the outputs were complete and fully consistent.

Outcomes of the 2001 Editing Process

8. Overall the 2001 Editing process was a success. However, the scanning and electronic capture techniques generated unexpected errors. For example:

Duplicate records: Where people had entered themselves more than once or had 'practised' completing the form on a page they did not intend using. Even where the 'extra' people had been crossed out, the recognition software still captured the data. A deterministic edit rule was applied which removed some 42,800 people, or 0.09 per cent of captured data.

Spurious People: The 1999 Dress Rehearsal identified the creation of spurious people originating from three causes:

- Erroneous marks on the form which the recognition software detected as a tick. These came from poorly printed constrained boxes or the recognition of 'marks' caused by dust on the scanner beds.
- Respondents crossing out pages that did not apply to the household; lines through constrained boxes were detected at recognition as a tick; and respondents writing 'N/A' or similar in the name field which caused them to be captured.
- Respondents turning over several pages at once, hence responding on pages that were intended for two different people.

A rule was applied to maximise the removal of spurious records and minimise the removal of real people. Some 3,297,800 person records (6.3 per cent) failed the rule and were removed from the data.

Minor Households: Households consisting of a single person aged under 16 years which came from:

- where date of birth was entered wrongly or recognised incorrectly; or
- where a Continuation form was not linked to the main form generating a new household.

The linkage problems also occurred where the unique identifier on one of the forms was either written or recognised incorrectly. As children were usually listed after adults in the household, Continuation forms mainly contained details of children.

III. CHALLENGES FOR 2011 QUESTIONNAIRE DESIGN

9. The structure and content of the 2011 Census paper questionnaire is not finalised but will be similar to 2001. However, offering every household in England and Wales the option to complete an on-line response raises a number of significant challenges for questionnaire design. ONS aim to design the Internet questionnaire in such a way as to minimise response bias. The paper and internet questionnaires will be identical in terms of question wording, response categories and instructions. However, the format of the on-line questions and questionnaire is likely to differ from the paper version. This difference may affect how people respond to the questions and could affect the accuracy of the responses between modes. Based on the experiences of other NSI's, we expect to observe

some, as yet unknown, degree of bias between the two collection modes. There are a number of factors that influence the respondent's experience of on-line data collection which include the overall look and feel of the interface, comparability with the paper questionnaire and the flow of questions.

10. To facilitate the use of on-line questionnaires, the paper version will have access codes pre-printed on the front page together with the website address where the Internet questionnaire can be located. The access codes will be unique, randomly generated numeric codes. On receiving the paper questionnaire, each household can choose to complete and return it by post or to complete and submit the online version. Where the response is by internet, the completed questionnaire will be transmitted to the processing centre, where it will be registered and the data integrated into the main paper flow of census returns.

11. We should bear in mind that respondents who choose internet data collection might be more experienced internet users, and have expectations about how the questionnaire should work based on experience with other web-based forms. Hence, there is a risk that, by adhering too closely to the existing paper questionnaire in an attempt to minimise the mode effect, we might actually induce or increase modal effects.

IV. CHALLENGES FOR DATA CAPTURE AND CODING

12. The introduction of an internet questionnaire raises a challenges for data capture and coding. For both the paper and on-line questionnaires, ONS has to carefully decide when and where to place the preliminary edits. For paper questionnaires, a single set of univariate edits will be applied at data capture to identify and correct for invalid responses. The question arises as to how these edits should be applied on-line. A basic on-line interface offers validation, automatic routing and on-line instructions or guidance. As the Internet questionnaire is developed we must consider the implications of increased functionality from both the ONS and the respondents' perspective. On balance we would expect to apply on-line functionality to improve data quality and reduce respondent burden. However, for some elements, increased functionality could actually increase respondent burden and introduce an additional element of response bias.

Scanning Error

13. Internet capture offers increased data quality by eradicating scanning error: if a response is 'typed' correctly it will be captured correctly. There will be no risk of spurious people resulting from dust on the scanner beds or poorly printed constrained boxes. Hence, there is a clear mode effect in terms of legibility of both tick and text responses between paper and on-line questionnaires. The possible between mode combinations of validity and legibility are shown in Table 1 overleaf. Both the paper and internet questionnaires can contain valid responses which are legible. Both can contain invalid responses which are legible. However, the on-line questionnaire cannot contain a valid response which is illegible i.e. due to poor handwriting or a poorly formed tick response both of which result in a missing value. Of course, a response to a paper questionnaire might be unreadable and also be wrong! For this aspect of data capture, the internet questionnaire generates clear quality gains.

Table 1: Tick and text response types by mode of collection

Response Type	Paper		Internet	
	Valid	Legible	Valid	Legible
1	✓	✓	✓	✓
2	✓	x	-	-
3	x	✓	x	✓
4	x	x	-	-

Automated Routing

14. A series of routing questions will be included in the 2011 paper questionnaire. Successful routing of any self-completion paper questionnaire depends on the provision of clear and concise instructions. Evaluation of the 2001 Census responses found that clear and concise instructions were present on the questionnaire. However, a number of respondents had difficulties understanding the requirements and did not follow the instructions. Hence, even with good questionnaire design, it is not possible to ensure that every respondent will read the questions and understand the routing.

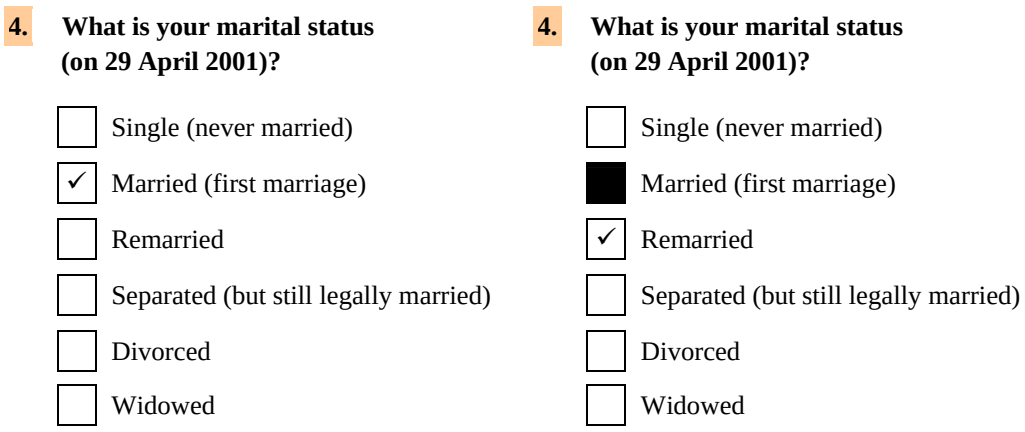
15. ONS are proposing that the design of the 2011 internet questionnaire should follow the paper version as closely as possible. Routing can be automated on-line. Questions that are 'not applicable', should not be presented to the respondent. However, respondents should be aware that they have skipped through the questions. The question numbering will be consistent for both modes, otherwise respondents might find it confusing to skip from question 15 to 20 without explanation. Well specified on-line routing has the potential to increase efficiency by reducing respondent burden, decrease overall completion time, reduce the volume of editing and increase data quality.

Radar Buttons and On-line Validation

16. The on-line interface provides the opportunity to apply a subset of the preliminary edits in real time. However, over editing could deter the respondents progress through the questionnaire, reducing the level of overall completion, whilst under editing could reduce quality at capture and inflate the volume of editing for later processing. For example, consider applying radar buttons for a multiple choice question which requires a single tick response. An experienced internet user might expect radar buttons to be activated and erroneously return a multi-ticked response if they were not. The implications of introducing, or not introducing, such functionality must be given careful consideration. In the example, efficiencies which could be achieved through the use of radar buttons are lost if they are not applied. There is a balance between adhering to procedures applied at paper capture and the incorporation of useful standard interface functionality.

17. To illustrate the value of radar buttons for a simple closed question, Figure 1 shows two valid responses to the Marital Status question from the 2001 Census. The question seeks a single ticked response as shown in the first example of Figure 1. A respondent making an erroneous response is instructed to black out the box and tick the 'correct' box (as per the second example in Figure 1). Both responses are valid and legible. However, evaluation of the 2001 Census found a large number of respondents erroneously multi-ticked without correcting the erroneous boxes as specified. Some respondents just put a line though the box which was subsequently scanned as a multi-ticked response. Identification and correction of multi-ticked responses places a significant load on paper capture and, for those unrepairable at capture, on the imputation. Hence applying radar buttons in an on-line questionnaire provides potential efficiency savings and increased data quality.

Figure 1: 2001 Census Marital Status Question – Examples of valid response
Left: single ticked response; Right: amended response



Since this is a simple closed question, some of the multi-ticked responses can be repaired at capture through preliminary edits of the type shown in Table 2. The remainder pass to imputation.

Table 2: Marital Status - Preliminary Edit for Paper Capture

QUESTION	RESPONSE	CODING RULES	OUTPUT CODE
What was your marital status on 29 April? Tick Boxes 1 Single 2 Married 3 Remarried 4 Separated 5 Divorced 6 Widowed	Single tick	Accept 1-6	1-6
	Multiple ticks	Two or three ticks accept one tick in priority order 4,3,5,6,2 Four or more ticks do not accept any tick. Code as W	2-6 W
	No ticks	Code as Y	Y

18. In addition to radar buttons, the use of online messages would prove useful to the editing process. Such messages could include those which:

- highlight item non-response or partial item non-response;
- highlight values outside of a pre-specified range for numeric responses; or
- simply ask for confirmation of implausible values.

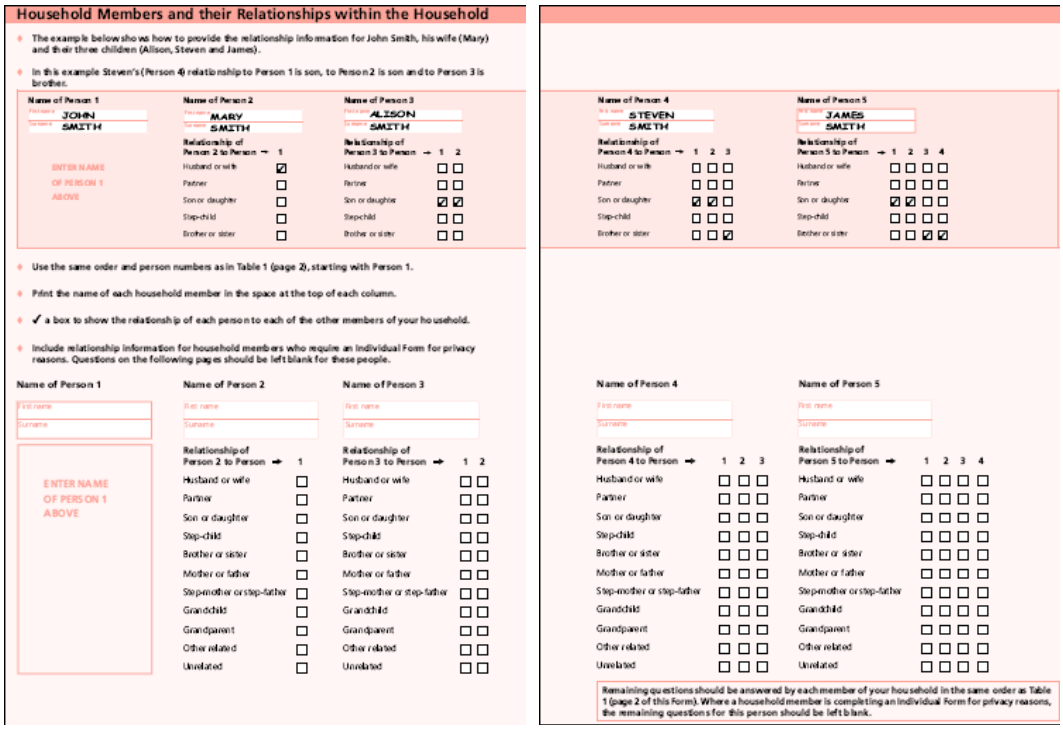
Thus, on-line messages could serve to highlight any missing, invalid or implausible responses to assist the respondent, who can then choose to either add or amend the information, or to continue to the next screen. Each time the respondent adds or amends any information the relevant edits should be invoked once again.

Personalising Questionnaires

19. The 2001 Census questionnaire asked respondents to record the names of all household members three times throughout the questionnaire: firstly, in the listing grid; then the relationship matrix; and finally the set questions relating to each person. The on-line interface offers efficiency savings and increased data quality by asking the respondent to record the names just once. For example, the respondent writes the names in the listing grid and the interface presents the names to the screen during later questions. The process personalises subsequent questions and removes impersonal terms, such as, ‘person 1’, ‘person 2’, and so on. Evaluation of 2001 responses found a large number records where household members were ordered differently in the listing grid, the relationship matrix and the person questions. Personalisation is a highly efficient technique with the potential to reduce respondent burden, reduce later complex editing algorithms (by ensuring a consistent ordering of individuals throughout the questionnaire) and improve data quality.

20. As an example, the 2001 Census questionnaire contained a relationship matrix which sought to identify family structure and hidden families by asking the relationships of all household members to one another. Figure 2, illustrates the 2001 Relationship Matrix. Briefly, the form-filler was asked to record the name of each household member and then to indicate the relationship of that person to other household members. The instructions were lengthy and some respondents found the layout confusing. Evaluation of 2001 data indicated that a number of respondents recorded the relationships the wrong way round for example, recorded Person 2 as the father of Person 1 when they were infact the son.

Figure 2: 2001 Census Relationship Matrix



21. The 2011 Census will include a relationship matrix of a similar format as 2001. In both response modes, the matrix will be preceded by a listing grid. The on-line interface has the potential to significantly improve data quality by writing the names recorded on the listing grid directly to the relationship matrix. The interface could offer questions such as, "How is Mary related to John?" as opposed to the bland, "Relationship of Person 2 to Person 1". A drop down menu could provide a list of the valid responses.

22. As a further example, information relating to the age of the population is a key requirement for the Census. Hence, the paper capture process seeks to achieve an almost 100% accuracy rate date of birth. The requirement for near perfect accuracy of date of birth also forms the cornerstone of the editing process. Personalisation techniques applied through the on-line interface could automate and present calculations relating to age. For example, where the form-filler records a date of birth as 21071961, in 2011 the derived age of 50 years and 4 months could be printed to the screen, with some appropriate text, to validate that the month and year of birth are accurate. These and similar validation techniques are available via on-line interfaces and will reduce the level of actual and perceived respondent burden, minimise the volume of edits in later processing, and improve data quality.

Drop Down Menus for Complex Coding

23. There are four basic steps in the coding process: achieving a response, constructing the classification ('coding frame' or 'nomenclature'), defining coding instructions (coding rules) and the coding method itself. The impact of complex coding on data quality can be measured in terms of: consistency, accuracy, throughput (number of keystrokes per minute but high number of keystrokes does not equal high quality); cost of the total coding operation and timeliness of delivery of the fully coded dataset. When complex coding is performed in an NSI, the major concerns are accuracy and consistency. Accuracy relates to allocating the most appropriate code to a description and is estimated by agreement with expert coders. Consistency is concerned with allocating the same code for items in the same category. So we need to ensure that individual coders assign the same code over time and that coders assign the same codes as each other. But this is dependent on the number of codes in the classification, the amount of factual interpretation needed and many different ways of describing the same instance. Further, during a census processing operation complex coding frames are extended during live running when a sufficient proportion of respondents report a new classification.

24. On the paper questionnaire questions such as Country of Birth, or Ethnicity are presented as a series multi-choice response boxes asking for a single tick or a tick and text response. For paper capture, textual responses are coded during processing using a comprehensive coding frame. The question is whether the coding frame should be embedded into the on-line interface and presented to the respondent as a drop down menu. For example, the 2001 Country of Birth Question was comprised of a series of tick boxes related to the countries of the United Kingdom (UK) and the Republic of Ireland. If the respondent was born overseas they were asked to tick the 'elsewhere' box and write their country of birth in a set of 20 constrained boxes. Figure 3 displays the 2001 Census Country of Birth question. The online interface could retain the 5 UK countries on the screen and offer the coding frame as a drop down menu allowing respondents who were born overseas to simply select their country of birth. However, the coding frame can never be exhaustive and must still contain the category 'elsewhere' or 'other' to allow capture of an element not contained within the frame.

Figure 3: 2001 Census Country of Birth Question

25. If embedding a complex coding frame into the interface was considered, then the format of the question could remain the same and the frame appear only when the 'elsewhere' box was ticked. There reasoning for this is that the majority of respondents select one of the first five boxes. Further, literature indicates issues of primacy and recency related to the ordering and number of categories presented to a respondent. On-line capture has the potential to generate a mode effect from embedding the full coding frame. There will be an additional affect arising from the presentation of the coding frame to the screen. Hence, research is needed to establish how best to format the list. A hierarchical presentation could be considered: first asking the respondent to select the continent, then to select the country within continent. Using this broad approach with questions where complex coding is only a component has the potential to achieve efficiency gains by reducing the volume of coding applied at later stages of processing. But the trade off may be a significant from loss of consistency and accuracy. It might prove more efficient overall for the respondent to continue with a simple write-in answer.

26. There are other questions which seek only a textual response and which are supported by a hierarchical complex coding frame. For example, consider the Standard Occupational Classification 2000 (SOC2000) which was developed by ONS, the Department of Trade and Industry and the University of Warwick. The classification is comprised of a hierarchical structure on 4 levels which can be aggregated to form a meaningful 1 digit summary. The 2001 Census SOC questions are shown above in Figure 4 followed by the coding frame in Table 4. This appears to be an example of where the question format should remain the same between modes. The alternative for internet would be to present the SOC in a hierarchical format in Question 27, instruct the respondent to select the correct category and then to 'type in' the response for Q28. The latter question only seeks further information to assist the coder to categorise the response to Q27. The mode effects and impact on data quality must be measured in terms of consistency and accuracy. Consistency is a concept that is conceptually difficult for on-line capture. Conventionally, consistency is allocating the same code for items in the same category. It is more likely that respondents will not navigate through the coding hierarchy successfully and will actually select the wrong category. This can only be estimated from pre-census small scale testing or a post censual quality survey.

Figure 4: 2001 Census - Standard Occupational Classification

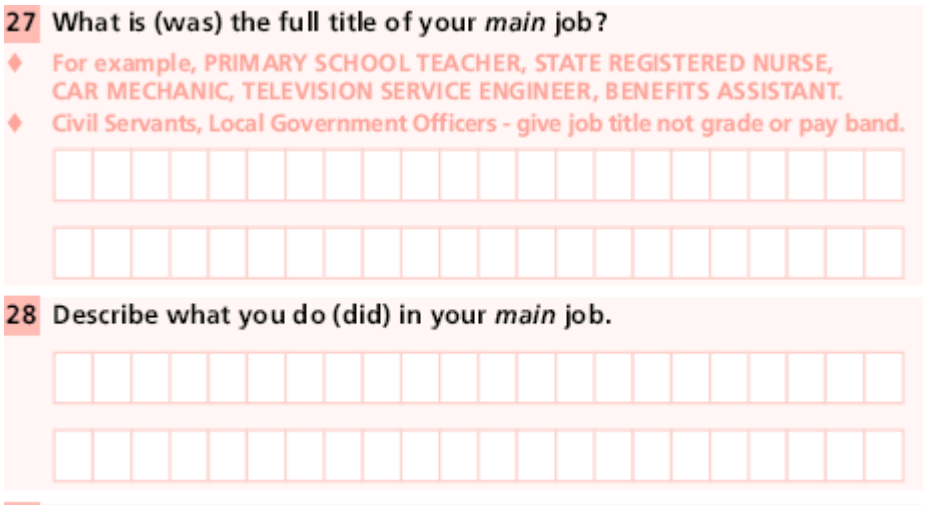


Table 4: Coding Frame for Standard Occupational Classification

Major Group	Occupation	Group		
		Sub-Major	Minor	Unit
1	Managers and Senior Officials	2	11	45
2	Professional	4	11	46
3	Associate Professional & Technical	5	17	53
4	Administrative & Secretarial	2	6	24
5	Skilled Trades	4	11	54
6	Personal Service	2	7	23
7	Sales & Customer Service	2	3	11
8	Process Plant & Machine Operatives	2	6	42
9	Elementary Occupations	2	9	35

27. Most coding operations are susceptible to errors. Coding error occurs due to coding decision errors, misapplication or insufficiency of the coding frame, and misapplication or insufficiency of the coding instructions. Coding errors can be random where the individual errors cancel out overall but if a correlated or systematic process is present bias will be introduced. Building complex coding frames into an on-line interface is high risk. As commented earlier, no matter how complete a coding frame appears it can never be considered exhaustive.

V. INTEGRATION OF PAPER AND INTERNET RESPONSES

28. The 2011 EW Census data capture and coding operation will be outsourced to a single supplier who will provide services for both the Internet and paper questionnaire. A key aim of the data capture and coding operation is to ensure that the maximum amount of valid, directly observed, information is supplied to ONS for further processing. The completed paper questionnaires will be mailed directly to a single data processing centre and registered against a master list of dwellings. The Internet responses will be transmitted to the data processing centre, where they will be registered and integrated into the flow of census returns.

29. The integration process is challenging and what is possible is dependent on the formatting of the Internet questionnaire and questions and the functionality, including editing, applied in the interface. All data captured from the internet will be transmitted to the processing centre and it is the format that will define the point of integration to the main data flow. At first thought, there are two broad options: if raw tick and text is received then integration would be immediately after scanning and prior to recognition; fully captured and simple coded data could be integrated immediately after

recognition and prior to edit. Similarly decisions about complex coding will be taken after deciding the options for the on-line interface.

30. Final decisions on whether to implement any on-line functionality, including editing, will have a direct impact on integration and may generate differences between modes of collection. The integration of records collected paper and on-line will be a highly tailored process. It will require co-operation between methodologists, subject matter experts and the external suppliers in designing the paper capture system and the on-line interface at all stages of the development cycle.

VI. CONCLUDING REMARKS

31. In respect of editing, ONS are seeking to develop a 2011 Census internet questionnaire that provides the optimum combination of response and quality, whilst also minimising bias. However, there is much work to do in order to understand the key drivers and levels of bias associated with applying multi-modes of collection in a census operation and to develop strategies to minimise it. There a clear need to understand the components of the user journey across both the internet and paper modes of collection.

32. Unfortunately, literature on the performance of the various editing approaches and their net effect on data quality is sparse. ONS is taking great care to ensure that the census does not suffer from over-editing, causing increased respondent burden, introduction of error and bias as edits are resolved, and increased processing time and cost. This paper has discussed a number of techniques and elements of functionality which could be applied on-line to reduce respondent burden, increase efficiency, improve data quality. The paper described how the use of on-line data collection provides the opportunity to capture higher quality data and eradicate scanning error. We have discussed simple and complex coding and concluded that as a minimum both the paper and on-line questionnaires must share the same basic coding frames. For internet capture, there is likely to be a mode effect from embedding and presenting the full coding frame. There will be an additional affect arising from how the full coding frame is presented. This broad approach to internet capture has the potential to reduce respondent burden, increase data quality and deliver efficiency gains by reducing the volume of subsequent editing.

33. The 2009 Census Dress Rehearsal provides the opportunity to test all prototype processing systems and operations, including internet collection, in preparation for the 2011 Census itself. The results of the Rehearsal will be scrutinised to identify and measure any mode effects introduced by Internet collection.

VII. REFERENCES

De Leeuw, E.D. (2005), “To Mix or Not to Mix Data Collection Modes in Surveys”. Journal of Official Statistics, Vol. 21, 233-255.

Dillman, D. and Christian, A.M. (2005) “Survey Mode as a Source of Instability in Response across Surveys”. Field Methods, Vol.17, 30-52.

ONS (2005) “Census 2001: Quality Report for England and Wales” HMSO. © Crown Copyright 2005.

Roy, L. and Laroche D. (2006) “The Internet Response Method: Impact on the Canadian Census of Population Data”. Journal of Survey Methodology.