

# Census 2001 Review and Evaluation

November 2002

## Edit and Imputation: Executive Summary

Content	Page
Project Objective.....	2
Background .....	2
Methodology.....	2
How well did it work.....	3
Lessons learned .....	3
Conclusion .....	3

ONS is carrying out a review and evaluation of the 2001 Census in England and Wales which will culminate in a Data Quality report and a General Report being published.

Plans for individual reports on specific aspects of the Census operation and a timetable for release have been published.

Each report is written in isolation and is subject to amendments as processing progresses and further information comes to light.

Reports will be released on the ONS website in the form of a high level Executive Summary and a more detailed Evaluation Report.

# Census 2001 Review and Evaluation

## Project Objective

Users of Census data asked the Census offices to provide output which is complete and consistent. They did not wish to fill in gaps in tables containing 'not known' responses by having to make their own estimates for missing values. The pattern of non-responses is often different from that of reported data, and without access to the individual records users would not be able to correct for such non-response bias or accurately estimate values for derived variables which were based on more than one item. There would also be a danger that different users would make estimates for the complete population in different ways, creating inconsistencies between their results. An Edit and Imputation strategy was therefore put in place with the aim of estimating for all missing data and resolving inconsistencies in responses for the people and households affected.

For the 2001 Census we aimed to follow these principles:

- All changes that were made would improve the quality of the data
- The number of changes to inconsistent data would be kept to a minimum
- As far as possible missing data would be imputed for all variables, so as to provide a complete and consistent database
- The system had to be relatively easy to develop and capable of processing large amounts of data automatically within short timescales.

## Background

An Edit and Donor Imputation System (EDIS) was devised for 2001 and applied to individual records once the data had been loaded by ONS. It was designed to fill in all the gaps in records for existing people and households, except for the voluntary question on religion.

The One Number Census (ONC) imputed for whole households and people who were missed from the Census at a later stage in processing. A separate evaluation report is being prepared on the ONC process.

## Methodology

EDIS contained five components:

- **Multi-tick rules** where more than one box was ticked but only one option was allowed.
- **Range checks** to prevent answers being outside an acceptable range.
- **Filter rules** to resolve some inconsistencies and to decide which fields should be set to 'No Code Required' where questions were answered but should not have been.
- A set of **Edit rules** to deal with missing items or responses which appeared to be in error or inconsistent when compared with other data. Edit either set a specific value or left it to imputation to determine a value.
- All items which were missing after the Edit stage were dealt with by the **Imputation** component which searched for a similar person or household (the donor) whose values would be copied into the records containing missingness (the recipient). A series of criteria were drawn up to determine what was meant by 'similar'. A suitable selection of variables (Primary Matching Variables) was defined to match on for each missing item. Rules were set up to cope with recipients with several missing items. There were also rules to ensure that each donor was not re-used too frequently.

The design of EDIS was such that it removed the bias that would otherwise have been created in the final statistics by missing responses. It is important to note that in some instances EDIS did not choose responses that were correct for the specific individual. However, it is the accuracy of the aggregate statistics that is critical, not the individual records.

# Census 2001 Review and Evaluation

---

## How well did it work

The three basic items Age, Sex and Marital Status, all had rates of missingness and imputation of well below 1 percent. The highest rate of non-response was 17.2 percent for the question on Professional Qualifications. There was an edit rule to set these cases to 'None' if the person had answered the previous question on Educational Qualifications. The second highest non-response rate was for Workplace size (13.9 percent). Where the worker was self-employed the size of workplace was inferred in Edit, otherwise it was imputed.

The eight most frequently executed edits accounted for 91 percent of the total.

For the household questions, non-response rates were between 2 percent and 4 percent except for number of rooms, which 5.4 percent of households failed to answer.

For some variables - such as Age and Sex - the proportion of imputed values assigned to each category closely followed the distribution of the reported data. For some other items there were significant differences, indicating that the people not replying to those questions were not typical of the remainder of the population.

Such biases were particularly noticeable for Activity Last Week, where fewer people were imputed as Working than among the people who responded to the question, and Retired and Student were imputed more frequently. Further analysis showed that people over 60 and those under 20 were much less likely to answer the relevant questions than those aged between 20 and 59, so that a method of estimating missing values of Activity Last Week which did not take account of age would have performed much less well.

With one exception, there was no means of checking the accuracy of imputation at an individual record level. However, for Sex the imputed values could in most cases be compared using people's names and a check for a 5 percent sample showed that 75 percent of values were correctly assigned. Among the remaining 25 percent the errors almost balanced out but a very slight excess of females was imputed. Note that edit rules were applied prior to this imputation stage so, for example, married couples of the same sex were not created.

## Lessons Learned

It was not possible to test the EDIS system on the data collected during the 1999 Dress Rehearsal, which meant that there was no opportunity to assess the efficacy of the imputation system or how well people would follow the filter instructions, which may have led to some amendments in the filter rules and edits.

Late changes to the questionnaire design had some impact on EDIS, in particular splitting the qualifications questions into two parts, which occurred after the 1999 Rehearsal. This change had a knock-on effect on to the question on Working last week, as a result of which some form-fillers answered No when they were actually in work. A change then had to be made to the filter rules for deriving Activity Last Week to correct for this misunderstanding.

## Conclusion

EDIS was successful in its main aim of providing a complete and consistent database of values for all people who completed Census returns. It did so efficiently and largely followed standard principles of making minimum changes to the data. There were complications in its development including late amendments, some of which could have been avoided with earlier access to live data and others which were due to changes after Rehearsal. However, these issues were identified at an early stage of Census processing and were quickly corrected.

# Census 2001 Review and Evaluation

Census Topics	Target Dates for Release
Legislation	Published
Non-Compliance (Executive Summary Only)	Published
Data Needs	Published
Geography	Published (Executive Summary)
Publicity	Published
Data Collection Development	Published
Data Collection Support	Published
Census Coverage Survey	Published
Processing	Published
Annex: Quality of Data Capture and Coding	Published
Downstream Processing	Published (Executive Summary)
Data Quality	
- Question non-response rates	Published
- Disclosure Control (Executive Summary only)	Published
- Data Validation (Executive Summary only)	Published
Edit & Imputation	Published
One Number Census	
- Quality Assurance	Published
- Lessons learnt (Executive Summary only)	Published
Output Policy	Published (Executive Summary)
Output Production	
- Part 1:Review of Output Released to date	Published (Executive Summary)
- Part 2:including Sample of Anonymised Records (SARs)/Origin Destination Matrices	Published
Census Access	Published
Programme Management	Published (Executive Summary)
Quality Report	Published
General Report	Published

Please note that the dates for release of individual evaluation reports noted above are target dates, and therefore subject to change. For the latest information please visit [www.statistics.gov.uk/census2001/reviewevaluation.asp](http://www.statistics.gov.uk/census2001/reviewevaluation.asp)