national STATISTICS

Census 2001 Review and Evaluation

February 2003

Edit and Imputation: Evaluation Report

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.

Content Page

Census Customer Services
ONS
Titchfield
Fareham
Hants PO15 5RR

Telephone: ++44 (0) 1329 813800
Fax: ++44 (0) 1329 813587
Minicom: ++44 (0) 1329 813669
E-mail: census.customerservices@ons.gov.uk
Website: www.statistics.gov.uk/census2001

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

For the 1991 Census, edit matrices had been developed for easy to code items to provide valid values to deal with inconsistencies. The most appropriate value was used, based on comparison with completed forms containing similar combinations.

Items could sometimes not be imputed in this way, because further inconsistencies arose. In these cases, and where there were missing values, an imputation system was put in place using a set of 50 tables which contained valid values for individual items. The tables were given an initial set of values based on earlier results, and as processing advanced they were updated with new values from wholly correct records. The most recent value, referenced by other characteristics of the person or household, was copied into the record requiring imputation known as the 'hot deck' method.

In developing a system for 2001, extensive tests were carried out on the use of neural networks which can detect complex relationships in data without the need for using complex modelling techniques. However, the testing failed to impute results which were consistent with the edit rules and neural networks did not perform as well as the 1991 hot deck method.

A donor edit and imputation system was also trialled, whereby all missing or inconsistent values on a record would be adopted from another similar individual (the donor). However, it was decided that setting specific values in the editing routines rather than basing them on similar donors would be more operationally efficient (although possibly less statistically accurate). Values would be set to missing for imputation if edit could not resolve an inconsistency. Thus an Edit and Donor Imputation System (EDIS) was devised for the 2001 Census. An extensive programme of specification and development of EDIS was undertaken which will be evaluated as part of the report on Downstream Processing.

EDIS was applied to individual records once the data had been loaded by ONS. It was designed to fill in almost all the gaps in records for existing people and households. A person was taken to exist if at least two of the name, date of birth and sex fields were completed. A process was applied prior to EDIS to remove any duplicate records, where someone had entered their data more than once or more than one form had been received for the same household.

The One Number Census process imputed for whole households and people who were missed from the Census. EDIS modified values for individuals on returned census forms.

Methodology

EDIS can be sub-divided into five elements:

Multi-tick rules dealt with cases where more than one box was ticked but only one option was allowed. In some cases there was a rule for selecting one tick.

If more than half the boxes were ticked or a set of priorities for accepting one tick could not sensibly be made, the answer was treated as missing, and a value was supplied at the imputation stage.

Range checks were applied to prevent answers being outside an acceptable range. These were set to missing for subsequent imputation. Examples were households with 0 or more than 99 rooms, or with more than 20 cars, people with a date of birth before 1891 or after Census Day, who last worked before 1941 or who worked more than 99 hours per week.

Filter rules were applied 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. For example, people under 16 or over 75 were not required to answer any of the employment questions. The variable Activity Last Week was also derived at this stage.

A set of Edit rules was applied to missing items or responses which appeared to be in error or inconsistent when compared with other data (such as married couples of the same sex, a child less than 13 years younger than its parents, or a married person under 16). These are known as hard checks.

In determining how to resolve such inconsistencies, the Fellegi/Holt principle of making the minimum number of changes was followed as far as possible. Thus if a person was under 16, married and had answered employment questions such as occupation, Age would be set to missing, since the inconsistency could be resolved with the least change by imputing a value for Age between 16 and 74.

Edit also identified unlikely, but not impossible responses. In some cases rules were applied to eliminate these: for example, a purpose-built flat was considered unlikely to have more than 10 rooms, and for reasons explained below the value was set to 'Missing' for imputation. In others no further action was taken, eg where people under 35 were retired from paid work. The number of these 'soft checks' was reported but the data were not changed as a result.

All items which were missing after the Edit stage were dealt with by the Imputation component, which is described below.

Imputation was applied when there was no answer on the Census form, it failed the multi-tick rules or was invalid, or the filter rules or Edit marked it for imputation to resolve an inconsistency.

The principle of a Donor Imputation System is to search for a single donor household to supply all the missing variables in a recipient household. Exceptions are imputation for postcode of usual address one year ago and of workplace, which were carried out at a later stage than imputation for other variables.

The search looked at all records in an Estimation Area, a group of contiguous Local Authority Districts of about 500,000 population. The method searched for a donor using up to five matching variables, which were determined by the fields requiring imputation on the recipient record. Values were copied over from the donor household to fill the missing values on the recipient record. Consistency checks were then applied and the donor rejected if any check failed.

Potential donor households were scored using a second set of matching variables relating to all people in the household. In addition, potential donors were penalised if they had been used before as a donor or if any of their fields had been edited or imputed. A record could not be used as a donor if any of the fields to be imputed were also missing on the donor. If potential donors still scored equally, the donor geographically closest to the recipient was chosen. However, to improve efficiency of the searching procedure, if a suitable donor was found who lived within 5,000 metres of the recipient, this person was accepted and no further search took place to find a closer donor.

The intention was to use joint imputation where possible, ie selecting a single donor household to impute for all the people with missing values in a recipient household so as to preserve the joint distributions between variables. If a suitable donor household could not be found for joint imputation, separate donors were sought to provide values for each person in the household requiring imputation, if necessary reducing the number of matching variables.

A fallback stage was also required as donor imputation failed to work for a few people. Most of these were imputed by testing possible values at random until one could be found which met the consistency criteria (a 'cold deck' approach). A small number of households could still not be completely resolved because of inconsistencies in age and relationships between people. As a final stage, ('son of fallback') if all else failed those containing up to eight people were completely replaced by synthetic households drawn at random from a set of the same household size, and households of nine or more people were corrected clerically.

The aim of imputation was to estimate the distribution of missing values accurately, so as to take account of any differences between the characteristics of respondents and non-respondents (non-response bias). It was not expected that the imputed values for every individual would be precisely accurate.

In comparison with 1991, EDIS was more comprehensive. It was applied to all variables, including qualifications, relationships, occupation, industry, hours worked, workplace address and means of transport to work, which were only analysed for a 10% sample of households and communals in 1991.

There was some manual intervention in the 1991 processing system, such as clerical checking of missing or inconsistent items which exceeded certain tolerances. EDIS was almost entirely automatic as clerical intervention was limited to households of more than eight people which failed the fallback stage.

Implementation

The EDIS system performed fully to specification, and ran well within the planned timetable. Imputation, which accounted for the bulk of the running time, was typically completed in 2 days for each of the 101 Estimation Areas in England and Wales.

In the analysis set out below, 'non-response' relates to failure to answer a question adequately, either because no response was supplied, or a value was out of range, inadequately described (in the case of occupation, industry or ethnic group), or multi-ticked. During Edit, some of these non-responses were set to a specific value. Edit also identified inconsistencies which led to a response being marked for imputation. The imputation rate may therefore be higher or lower than the rate of non-response.

People and households imputed by the One Number Census process are not included in these analyses.

Edit

A total of 13.7 million (m) edits were carried out on the data for 11.8m people. The base population for EDIS was 49.4m people in England and Wales, including some 0.6m students living away from home during term-time for whom only a few demographic and relationship questions applied at their home address. The eight most frequently executed edits accounted for 91% of the total. These were:

4.50m	Professional qualifications set to None where missing but educational qualifications was answered
2.29m	Carer set to No where missing unless Activity Last Week was also missing
1.66m	Workplace size set to 1-9 where person was self-employed
1.08m	Travel to work set to "work mainly at/from home" where workplace address was "mainly work at/from home"
1.03m	Supervisor set to No if missing, unless occupation was also missing
1.01m	Health set to Good if missing, unless Activity Last Week was also missing
0.59m	Professional qualifications set to missing if answered but educational qualifications was missing
0.40m	Missing Country of birth set to that of either siblings, parents or other related people in the household who have the same Country of birth

Imputation

One or more items needed to be imputed for 13.8m people - that is 28.0% of the population who returned Census forms. Of these, 4.7m were dealt with by joint imputation. 10.0m were imputed using individual imputation, including all those in single person

households. 9.8m of the individual imputed cases used a donor household of the same size as the recipient's and the remaining 0.2m a household of different size. 0.4m people required imputation using the cold deck fallback method. Over 1m people had some items imputed by one method and some by another, hence there is some double-counting.

23.4% of the population were used once as donors, 2.1% twice and 0.1% three or more times.

For household variables, 2.5m needed imputation, 11% of all households. 0.08m were dealt with by fallback and the remainder by joint imputation. Almost all the donor households for joint imputation were used once each.

Person variables

	Total (including imputed)	Non-response	Imputed	Non-response	Imputed
	000s	000s	000s	%	%
Age	49,359	262	278	*0.53*	*0.56*
Sex	49,359	199	219	*0.40*	*0.44*
Marital status	49,359	372	158	*0.76*	*0.32*
Student flag	49,359	622	641	*1.26*	*1.30*
Country of birth	48,848	1,211	829	*2.48*	*1.70*
Ethnic group	48,848	1,405	1,421	*2.88*	*2.91*
Welsh language	2,754	153	153	*5.54*	*5.57*
Religion	48,848	3,721	-	*7.62*	-
Health	48,848	1,525	531	*3.12*	*1.09*
Carer	48,848	2,967	693	*6.07*	*1.42*
Long-term illness	48,848	1,899	1,915	*3.89*	*3.92*
Address one year ago	48,848	2,198	2,213	*4.50*	*4.53*
Educational qualifications	35,367	2,187	-	*6.18*	-
Professional qualifications	35,367	6,094	-	*17.23*	-
Highest qualification	35,367	-	2,150	-	*6.09*
Working last week	35,367	737	-	*2.08*	-
Activity last week	35,367	-	1,301	-	*3.69*
Employment status	33,686	2,205	2,058	*6.55*	*6.14*
Workplace size	33,686	4,689	3,067	*13.92*	*9.15*
Supervisor	33,686	2,294	1,119	*6.81*	*3.34*
Occupation - currently working	21,741	694	759	*3.19*	*3.48*
Occupation - all ever worked	29,335	4,051	4,051	*13.81*	*13.81*
Industry - currently working	21,741	1,702	1,777	*7.83*	*8.15*
Industry - all ever worked	29,335	5,400	5,400	*18.41*	*18.41*
Workplace address	22,396	1,744	1,426	*7.79*	*6.42*
Method of travel	22,533	1,410	1,127	*6.26*	*5.07*
Hours worked	22,533	1,804	1,506	*8.00*	*6.77*
Relationship to Person 1	28,065	971	1,326	*3.46*	*4.73*

Note: The 'Total' column refers to the number of people in scope for the question, ie:

- Age, Sex, Marital status, Student flag: All people plus students who were counted at both their home address and term-time address in England and Wales
- Country of birth, Ethnic group, Religion, Health, Carer, Long-term illness, Address one year ago: All people (students counted at term-time address only)
- Welsh language: All people living in Wales
- Qualifications, Working last week: All people aged between 16 and 74
- Employment status, Company size, Supervisor: All people aged between 16 and 74 who had ever worked
- Workplace address, Method of travel, Hours worked: All people aged between 16 and 74 who were working in the week before Census day
- Relationship to Person 1: All people in households plus students also counted at home address less those who were entered as Person 1 on census form

Age

Age was not reported or was out of range (born after Census day or more than 110 years old) for 240,000 people. It was set to missing for a further 23,000 on grounds of inconsistency, mainly because people who were not single and who had answered three or more employment questions had their age captured as under 16.

	Imputed		Total (including imputed)	
Age group	000s	%	000s	%
0 - 4	15	*5.6*	2,800	*5.7*
5 - 9	12	*4.4*	3,066	*6.2*
10 - 14	10	*3.7*	3,233	*6.5*
15 - 19	26	*9.3*	3,143	*6.4*
20 - 24	23	*8.4*	3,032	*6.1*
25 - 29	21	*7.4*	3,061	*6.2*
30 - 34	22	*8.0*	3,655	*7.4*
35 - 39	20	*7.3*	3,819	*7.7*
40-44	19	*7.0*	3,461	*7.0*
45-49	18	*6.3*	3,157	*6.4*
50-54	21	*7.6*	3,470	*7.0*
55-59	16	*5.8*	2,876	*5.8*
60-64	16	*5.8*	2,476	*5.0*
65-69	12	*4.3*	2,238	*4.5*
70-74	11	*4.0*	2,029	*4.1*
75-79	6	*2.0*	1,717	*3.5*
80-84	4	*1.6*	1,146	*2.3*
85 & over	4	*1.6*	984	*2.0*

The distribution of imputed ages followed that of the remainder of the population except for a shortfall among the 0, 6-15 and 76-80 age groups. This is primarily because some people were imputed as aged between 16 and 74 who may have been outside this age range because some employment questions had been answered. The shortfall in babies under 1 year old occurred where their address one year ago had not been stated as 'no usual address'. The effect in an area of 100,000 population would typically be that 2 or 3 under 1's would have been imputed as over 1.

Sex

Sex was missing for 185,000 people and multi-ticked for 14,000, 0.4% of the population in total. There were no edit actions which directly affected this question: if a husband and wife, or the parents of a child, were of the same sex the relevant relationships were imputed. A further 20,000 had values imputed by 'son of fallback'.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Female	113	*51.7*	25,473	*51.6*
Male	106	*48.3*	23,887	*48.4*

The sexes were imputed in the ratio of 51:49 in favour of females, very similar to the proportions among the remainder of the population. The accuracy of imputations was assessed by comparing the imputed values with people's names in a sample of areas. This showed that 75% of imputations were correct. Among the incorrect values there was a very slight bias towards imputing females. The net effect would be to count four people out of every 100,000 as female rather than male.

Marital Status

There were 373,000 missing or multi-ticked cases for marital status, representing 0.8% of the population. 232,000 of these were children under 16 who were set to Single in edit. A further 6,000 under 16s had marital status changed to Single. Imputation was applied to the remainder. Married and Re-married were less likely to be imputed than among the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Single	74	*46.7*	21,440	*43.4*
Married	41	*26.0*	17,517	*35.5*
Re-married	9	*5.5*	2,975	*6.0*
Separated	4	*2.6*	895	*1.8*
Divorced	13	*8.2*	3,186	*6.5*
Widowed	17	*11.0*	3,346	*6.8*

Census 2001 Review and Evaluation

Student

Question 5 on the person schedule asked whether a person was a schoolchild or student in full-time education. 1.3% of people failed to answer or multi-ticked the question, of whom 13% were imputed as students compared with 21% in the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Student	85	*13.3*	10,479	*21.2*
Not a student	556	*86.7*	38,881	*78.8*

Country of Birth

Country of birth was omitted by 2.5% of people. Of these, 88% were imputed as born in the United Kingdom, compared to 92% in the remainder of the population. People born in Africa, Asia and North America were imputed in higher proportion than the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
UK	727	*87.7*	44,769	91.7
Republic of Ireland	9	*1.1*	440	*0.9*
Europe	18	*2.2*	930	*1.9*
Africa	18	*2.1*	674	*1.4*
Asia	38	*4.6*	1,384	*2.8*
North America	14	*1.7*	407	*0.8*
South America	2	*0.2*	65	*0.1*
Oceania and other	3	*0.4*	176	*0.4*

Ethnic Group

The non-response rate for ethnic group was 2.9%. 89% of these were imputed as White compared with 92% in the remainder of the population. There were higher proportions of imputed people in the Mixed, Asian and Black groups.

	Imputed		Total (including imputed)	
	000s	%	000s	%
White	1,260	*88.7*	45,065	92.3
Mixed	24	*1.7*	605	*1.2*
Asian	80	*5.6*	1,925	*3.9*
Black	43	*3.0*	868	*1.8*
Chinese and other	13	*0.9*	382	*0.8*

Welsh Language

The question asking whether people could understand spoken Welsh, or speak, read or write the language, was asked of all people living in Wales. There was a 5.5% non-response rate. No knowledge of Welsh was imputed slightly more often than for the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
No knowledge	112	*73.6*	1,974	72.4
Understand only	6	*4.2*	133	*4.9*
Speak only	1	*1.0*	26	*1.0*
Read only	1	*0.9*	28	*1.0*
Write only	0	*0.2*	4	*0.1*
Understand, speak, read and write	20	*13.0*	382	*14.0*
Other combinations	11	*7.2*	179	*6.6*

Census 2001 Review and Evaluation

Religion

As the question on religion was voluntary, non-responses were not imputed but will appear in tables as 'not stated'. The national non-response rate was 7.6%.

General Health

This question asked whether over the last twelve months a person's health had on the whole been good, fairly good or not good. The non-response rate was 3.1%, but an edit rule set the value to good unless Activity Last Week was also missing. This reduced the number requiring imputation to 1.1%. Among these people, Fairly Good and Not Good were imputed slightly more frequently than in the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Good	349	*65.6*	33,348	*68.3*
Fairly good	130	*24.5*	10,948	*22.4*
Not good	52	*9.8*	4,549	*9.3*

Carer

Question 12 referred to voluntary help or support given to family members, friends or neighbours. The rate of non-response was 6.1%. Missing values were set to No by an edit rule unless Activity Last Week was also missing, and children under 5 were also assumed to not be providing care. Of the remaining 1.3% of the population, 11% were imputed as Carers in comparison to 10% among the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Not a carer	615	*88.6*	43,833	*89.7*
1 to 19 hours	52	*7.4*	3,417	*7.0*
20 to 49 hours	9	*1.3*	547	*1.1*
50+ hours	18	*2.6*	1,048	*2.1*

Long-term Illness

There was a 3.9% non-response rate to this question, which asked about any long-term illness, health problem or disability which limited the person's daily activities or the work they could do. 22% of these were imputed as having such a condition in comparison with 18% among the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Long-term illness	422	*22.1*	9,042	*18.5*
No long-term illness	1,492	*77.9*	39,803	*81.5*

Address One Year Ago

This question had a non-response rate of 4.5%. No usual address was imputed more often than among the remainder of the population, mainly because there was a high rate of non-response for children under 1.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Address shown on front of form	1,722	*77.8*	40,839	*83.6*
No usual address one year ago	145	*6.6*	668	*1.4*
Same as Person 1	149	*6.7*	3,098	*6.3*
Elsewhere	197	*8.9*	4,241	*8.7*

Census 2001 Review and Evaluation

Qualifications

This topic was covered by two questions, on educational and professional qualifications, which had non-response rates of 6.2% and 17.2% respectively. Where missing, professional qualifications was set to None by an edit rule if the educational qualifications was answered. Professional qualifications was set to missing if educational qualifications was not answered. Taking the responses to the two questions together, a new variable called highest qualification was derived. After applying the edit rules, 6.1% of people needed to have highest qualification imputed. People with imputed values were more likely to have no qualifications (Level 0) than the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Level 0	846	*39.3*	10,387	*29.4*
Level 1	297	*13.8*	5,866	*16.6*
Level 2	366	*17.0*	6,843	*19.4*
Level 3	142	*6.6*	2,841	*8.0*
Level 4	336	*15.6*	6,885	*19.5*
Level 5	163	*7.6*	2,495	*7.1*

Activity Last Week

This variable shows whether a person was working in the week prior to Census day, and if not whether they were looking for work, waiting to start a job, retired, student, looking after home/family, permanently sick or disabled, or otherwise economically inactive. This information is derived from Questions 18 to 22 on the Census form for people aged 16 to 74.

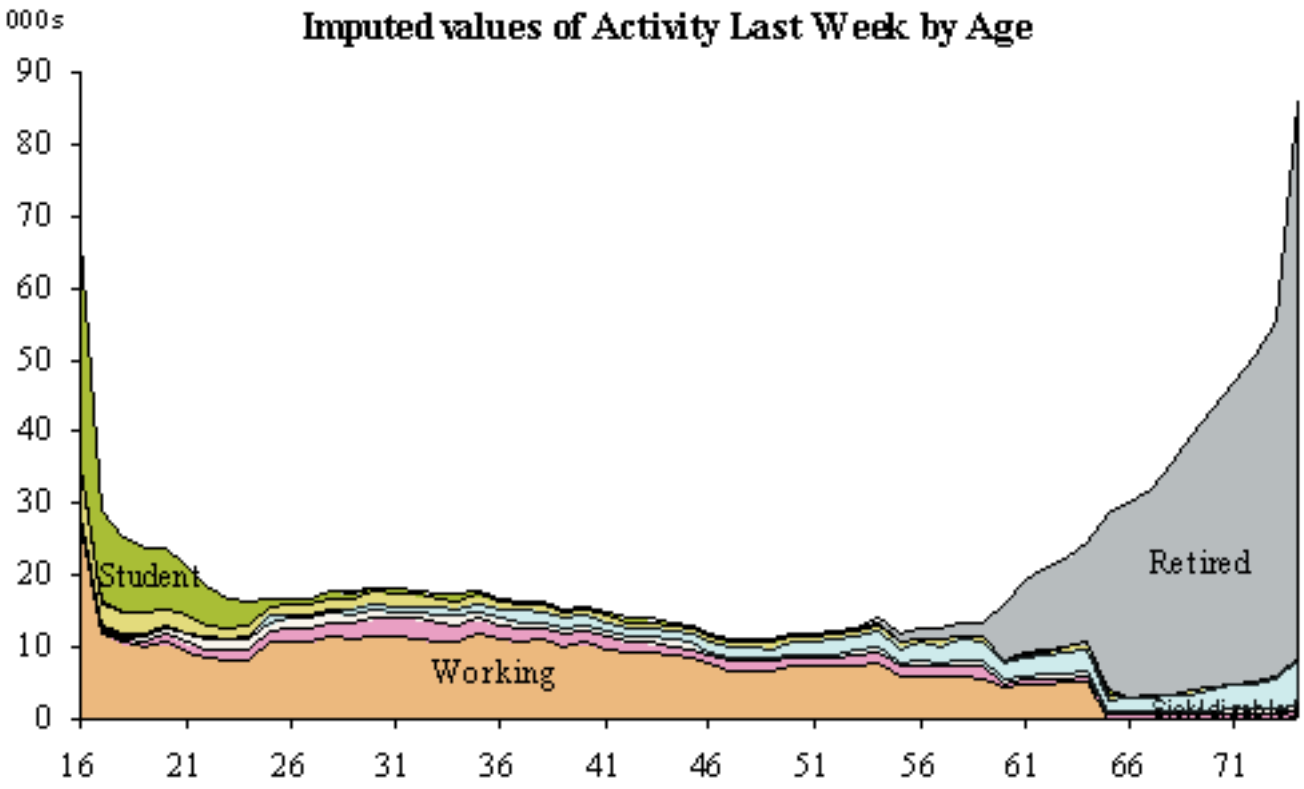
Problems were found with the pattern of responses to these and other employment questions which was caused by the format of Question 18 (Last week, were you doing any work). Some people ticked No or multi-ticked this question, but then went on to give details of their present job in answer to Questions 32 to 36. The filter rules were amended to accommodate this pattern so that they were treated as working.

Non-response to working last week was 2.1%. The value was changed in certain cases depending on the pattern of responses to looking for work etc (questions 19-22), ever worked and year last worked (question 23), details of current or last job at questions 25-30 and current job at questions 32-35.

	Imputed		Total (including imputed)	
	000s	**%**	**000s**	**%**
Working	438	*33.7*	22,228	62.9
Looking for work	67	*5.1*	1,213	3.4
Waiting to start a new job	1	*0.1*	25	*0.1*
Retired	470	*36.1*	5,002	*14.2*
Full time education	107	*8.2*	1,545	*4.4*
Looking after home/family	76	*5.8*	2,290	*6.5*
Permanently sick/disabled	96	*7.4*	1,947	*5.5*
Other economically inactive	46	*3.6*	1,068	*3.0*

In total, 3.7% of Activity Last Week values were imputed. These were biased towards looking for work and most of the economically inactive categories, especially retired and students. Only 34% were imputed as working compared with 64% in the remainder of the population aged 16-74. The graph below shows that it was generally people at the extremes of the age range who failed to respond to these questions, which explains the preponderance of retired people and students among the imputed values.

Census 2001 Review and Evaluation



Employment Status

Question 25 asked whether each person who had ever worked was an employee, or self-employed with or without employees in their current or last job. Non-responses and multi-ticks amounted to 6.5% of those who should have answered the question. These all went through imputation, and 'Employee' was imputed more frequently than among the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Employee	1,918	*93.2*	29,552	*88.1*
Self employed:				
With employees	75	*3.6*	1,472	*4.4*
Without employees	65	*3.2*	2,504	*7.5*

Size of Workplace

The non-response rate for this question was 13.9%. An edit rule was applied to set the number of workers to 1-9 where a person was self-employed without employees. This left 6.5% to be imputed, of whom slightly fewer were set to 1-9 workers than among the remainder of the population, and slightly more in the 10-24 and 25-499 ranges.

	Imputed		Total (including imputed)	
	000s	%	000s	%
1 to 9 employees	*737*	*24.0*	9,204	*27.5*
10 to 24	521	*17.0*	5,108	*15.2*
25 to 499	1,214	*39.6*	12,753	*38.0*
500 or more	594	*19.4*	6,463	*19.3*

Occupation and Industry

The non-response rate for occupation was 3.2% among currently working people, including 0.7% inadequately described responses. When all people who had ever worked are considered, non-response rose to 13.1%. The imputed population was slightly biased towards people in major groups 4 (administrative and secretarial), 7 (sales and customer services), 8 (process, plant and machine operatives) and 9 (elementary occupations). Occupation groups 2 (professional) and 3 (associate professional and technical occupations) were under-represented.

	Currently working people				All ever worked people			
wSOC2000			Total (including				Total (including	
	Imputed		imputed)		Imputed		imputed)	
Major group	000s	%	000s	%	000s	%	000s	%
1	114	*15.0*	3,311	*15.2*	459	*11.3*	4,034	*13.8*
2	78	*10.2*	2,410	*11.1*	320	*7.9*	2,943	*10.0*
3	96	*12.7*	2,976	*13.7*	402	*9.9*	3,674	*12.5*
4	104	*13.8*	2,915	13.4	622	*15.4*	4,010	*13.7*
5	86	*11.3*	2,566	*11.8*	363	*9.0*	3,235	*11.0*
6	46	*6.1*	1,517	*7.0*	333	*8.2*	2,179	*7.4*
7	62	*8.2*	1,660	*7.6*	424	*10.5*	2,538	*8.7*
8	72	*9.5*	1,868	*8.6*	412	*10.2*	2,622	*8.9*
9	101	*13.3*	2,576	*11.8*	717	*17.7*	4,100	*14.0*

Note: Data analysed according to the major groups of the Standard Occupation Classification 2000 (SOC2000)

A similar pattern can be found in non-response to the question on industry. Non-response was 7.8% among current workers, including 0.6% inadequately described, but reached 17.9% taking into account all people who have worked. Imputation created more people working in sections A (agriculture), F (construction) and O (social and personal services) and fewer in D (manufacturing), J (banking, finance, insurance), L (public administration) and M (education).

	Currently working people				All ever worked people			
SIC(92)			Total (including				Total (including	
	Imputed		imputed)		Imputed		imputed)	
Section	000s	%	000s	%	000s	%	000s	%
A	62	*3.5*	341	*2.0*	104	*1.9*	431	*1.5*
C	4	*0.2*	56	*0.3*	15	*0.3*	80	*0.3*
D	247	*14.0*	3,297	*19.8*	825	*15.3*	4,474	*15.3*
E	9	*0.5*	160	1.0	33	*0.6*	216	*0.7*
F	194	*11.0*	1,496	*9.0*	389	*7.2*	1,899	*6.5*
G	399	*22.6*	4,687	*28.2*	1,329	*24.6*	6,748	*23.0*
I	126	*7.1*	1,519	*5.1*	350	*6.5*	1,964	*6.7*
J	273	*15.5*	3,799	*13.4*	777	*14.4*	4,866	16.6
L	68	*3.8*	1,249	*4.0*	239	*4.4*	1,575	*5.4*
M	272	*15.4*	4,083	*13.3*	1,041	*19.3*	5,578	*19.0*
O	113	*6.4*	1,111	*3.9*	297	*5.5*	1,503	*5.1*

Note: Data analysed according to the sections of the UK Standard Industrial Classification of Economic Activities 1992 (SIC(92)). Sections A and B were combined for imputation purposes, as were sections G and H, sections J and K, sections M and N, and sections O, P and Q.

It should be noted that the full codes were imputed for missing occupation and industry data. However, the primary matching variables for these fields were defined at the major group level. Thus if industry was reported but occupation was missing, a donor would have been sought within the same major industry group, and that person's occupation copied into the recipient's record. In some cases an unlikely occupation/industry combination may have been created at the individual code level.

Supervisor

Question 29 asked whether people supervised any other employees in their current or last job. The non-response rate was 6.8%. An edit rule set missing answers to No unless occupation was also missing. This accounted for about half the non-response. Of the remainder, 25% were imputed as supervisors compared with 30% among the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Supervisor	284	*25.4*	10,014	*29.9*
Not a supervisor	835	*74.6*	23,515	*70.1*

Workplace Address

There was a 7.8% rate of non-response to this question, but some values could be deduced from the answers to method of travel to work. This left 6.4% to be imputed. Of these, fewer were imputed as working at or from home than amongst the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Mainly at/ from home	49	*3.5*	2,061	*9.3*
Offshore installation	1	*0.1*	14	*0.1*
No fixed place	89	*6.2*	977	*4.4*
Address below	1,287	*90.2*	19,179	*86.3*

Method of Travel to Work

This question was asked only of currently working people. Non-response was 6.3%, which was reduced to 5.0% by a set of edits. The imputed values were biased towards public transport users and those travelling by foot and away from working at/from home or driving a car or van.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Mainly at/from home	49	*4.4*	2,061	9.3
Underground, Metro, light rail or tram	40	*3.5*	576	*2.6*
Train	45	*4.0*	876	*3.9*
Bus, minibus or coach	102	*9.1*	1,596	*7.2*
Motorcycle, scooter or moped	12	*1.1*	243	*1.1*
Driving a car or van	613	*54.4*	12,449	*56.0*
Passenger in a car or van	81	*7.1*	1,405	*6.3*
Taxi	8	*0.7*	113	*0.5*
Bicycle	32	*2.9*	608	*2.7*
On foot	138	*12.2*	2,200	*9.9*
Other	7	*0.6*	103	*0.5*

Hours worked

The non-response rate was 8.0%, and imputation favoured the 0-19 hours per week range compared to the pattern among the remainder of the population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
0 to 19 hours	196	*13.0*	2,679	*12.1*
20 to 29 hours	146	*9.7*	2,193	*9.9*
30 to 39 hours	451	*30.0*	6,798	*30.6*
40 to 49 hours	476	*31.6*	7,012	*31.5*
50 hours or more	237	*15.8*	3,547	*16.0*

Census 2001 Review and Evaluation

Household Variables

	Total (including imputed)	Non-response	Imputed	Non-response	Imputed
	000s	000s	000s	%	%
Accommodation type	22,305	671	671	*3.01*	*3.01*
Self-contained	22,305	870	870	*3.90*	*3.90*
Number of rooms	20,542	1,117	1,116	*5.44*	*5.21*
Bath/shower and toilet	20,542	503	503	*2.45*	*2.35*
Lowest floor level	22,305	897	919	*4.02*	*4.12*
Central heating	20,542	539	442	*2.62*	*2.17*
Number of cars	20,542	669	554	*3.26*	*2.72*
Tenure	20.542	797	685	*3.88*	*3.36*
Landlord	6,582	-	175	-	*2.94*

Note: The 'Total' column refers to the number of households in scope for the question, ie:
- Accommodation type, self-contained, lowest floor level: All household spaces whether occupied or not (including enumerators' responses on dummy forms where no Census return was made)
- Number of rooms, bath/shower and toilet, Central heating, Number of cars, Tenure: All occupied household spaces, ie households with at least one usual resident
- Landlord: All occupied household spaces where Tenure was renting or rent free

Accommodation Type

There was a 3.0% non-response rate for this question, which was asked of all households. Imputed values were more likely to be a purpose-built flat, part of a converted or shared house, or a commercial building, and less likely to be a detached or semi-detached house.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Detached house	148	*22.1*	5,112	*22.9*
Semi-detached house	171	*25.6*	7,067	*31.7*
Terraced house	177	*26.3*	5,818	*26.1*
Purpose-built flat	111	*16.6*	3,023	*13.6*
Part of converted or shared house	47	*7.0*	936	*4.2*
Commercial building	13	*1.9*	256	*1.1*
Caravan, mobile or temporary	4	*0.6*	93	*0.4*

Self-contained

This question had a non-response rate of 3.9%. Of imputed households, 1.5% were given not self-contained status compared with 1.1% among the remainder of the household population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Self-contained	858	*98.5*	22,055	*98.9*
Not self-contained	13	*1.5*	250	*1.1*

Number of Rooms

Question H3 provided two boxes for the number of rooms occupied by a household, so that any value from 1 to 99 could be entered. Early analysis of processed data showed that there were some problems which needed to be addressed:

- A zero entered into the left hand box was sometimes interpreted by OCR as a 6, creating values from 61 to 69 instead of 1 to 9.

- A diagonal slash entered into the left hand box was sometimes mistaken for a 1, creating values from 11 to 19.

- If the form-filler attempted to make a change by crossing out and writing a different figure in the other box, both figures might be recognised by OCR or the number could be duplicated in clerical processing. Thus where a value of 3 was changed to 4, the number of rooms might have been interpreted as 33, 34, 43 or 44.

After carrying out an analysis of households with more than 10 rooms, rules were put in place to set values to missing where they were greater than a number which depended on accommodation type. Number of rooms was subsequently imputed. No limit was applied to detached houses.

Imputation was slightly more likely to set a value of 3 or 4 rooms, and less likely to impute 5 or more rooms, compared with the remainder of the household population.

No of Rooms	Imputed		Total (including imputed)	
	000s	%	000s	%
1	9	*0.8*	180	*0.8*
2	27	*2.5*	518	*2.4*
3	106	*9.5*	1,907	*8.9*
4	233	*20.9*	4,235	*19.8*
5	295	*26.4*	5,831	*27.2*
6	229	*20.5*	4,462	*20.8*
7	100	*9.0*	2,010	*9.4*
8	58	*5.2*	1,164	*5.4*
9	29	*2.6*	591	*2.8*
10	15	*1.3*	288	*1.3*
11 or more	14	*1.2*	256	*1.2*

Bath/shower and toilet

Question H4 asked whether a bath/shower and toilet was available for use only by the household. There was a non-response rate of 2.5%, and slightly more households were imputed as lacking sole use than among the remainder of the household population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Exclusive use	500	*99.4*	21,341	*99.5*
Lacking or shared use	3	*0.6*	101	*0.5*

Lowest floor level

4.0% of households failed to answer this question. Fewer were imputed as having ground floor as their lowest level of accommodation than the remainder of the household population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Basement or semi-basement	33	*3.6*	600	*2.7*
Ground floor	769	*83.6*	19,107	*85.7*
First floor	72	*7.8*	1,677	*7.5*
Second floor	24	*2.6*	506	*2.3*
Third or fourth floor	14	*1.5*	265	*1.2*
Fifth floor or higher	8	*0.9*	150	*0.7*

Central heating

This question had a non-response rate of 2.6%. Non-respondents were slightly more likely to lack central heating than for the remainder of the household population.

Census 2001 Review and Evaluation

	Imputed		Total (including imputed)	
	000s	%	000s	%
Has central heating	398	*90.0*	19,592	*91.6*
Lacks central heating	44	*10.0*	1,804	*8.4*

Number of cars

There was a 3.3% non-response to this question. 35% of these households were imputed as having no cars compared with 26% for the remainder of the household population.

	Imputed		Total (including imputed)	
	000s	%	000s	%
No cars	192	*34.6*	5,694	*26.6*
One car	247	*44.6*	9,372	*43.8*
Two cars	91	*16.5*	5,062	*23.7*
Three cars	18	*3.3*	970	*4.5*
Four cars	4	*0.7*	225	*1.0*
Five or more	2	*0.3*	73	*0.3*

Tenure and Landlord

Non-response to these questions was 3.9% for tenure and 2.9% for landlord. Those not answering were more likely to be renting and less likely to be outright owners than in the remainder of the population. Among tenants, there was little bias towards any type of landlord among the imputed group.

	Imputed		Total (including imputed)	
	000s	%	000s	%
Tenure				
Owns outright	165	*24.1*	6,343	*29.6*
Owns with mortgage or loan	197	*28.8*	8,342	*39.0*
Part rent part mortgage	3	*0.4*	130	*0.6*
Rents	296	*43.2*	6,141	*28.7*
Lives rent free	24	*3.5*	441	*2.1*
Landlord				
Council (Local Authority)	79	*44.8*	2,984	*45.3*
Housing Association, Charitable Trust etc	34	*19.5*	1,312	*19.9*
Private landlord or letting agency	47	*26.7*	1,837	*27.9*
Employer	5	*2.7*	124	*1.9*
Relative or friend	8	*4.7*	230	*3.5*
Other	3	*1.7*	95	*1.4*

Comparison with 1991

In general, the biases found in the imputed values for the person and household variables were in the same direction as those present in the 1991 Census data, but were less marked. For example, 52% of those imputed in 1991 for marital status were assigned as Single compared with 41% in the Census population. In 2001 the corresponding proportions were 49% and 44%.

53% of non-respondents were imputed as having no car in 1991, considerably higher than the 32% among reporting households. In 2001, when the non-response rate had risen from 1.0% to 3.3%, the gap had narrowed to 34% among imputed households and 27% in those who responded.

Lessons learned

Although EDIS in general performed well and within planned running times, some aspects might have worked better if there had been more opportunity for testing.

It did not prove possible to carry out a full run-through of the EDIS system on the data collected during the 1999 Dress Rehearsal. This would have afforded the opportunity of checking whether ideas which appeared sensible in theory would stand the test of practical application on live data. Rehearsal data needed to be delivered more rapidly, at least on a small scale, or the Rehearsal itself should have been brought forward so that it took place more than two years before Census day.

Inability to test the system meant that assumptions about how well the public would follow instructions, for example on answering or skipping certain questions, proved to be not entirely valid. This impacted most noticeably on imputation of age, although the extent of systematic error turned out to be very slight.

Within EDIS, a number of assumptions were based on age being correct rather than other items. However, year of birth was occasionally mis-stated, not scanned correctly or given a wrong value during processing. Particularly when there was an error in the next to last digit of the year, EDIS may have imputed for a range of items where no value was needed, or conversely set reported data to 'no code required'. Further checks could have been tested if more time had been available to investigate real data from the Rehearsal: for example, identifying circumstances where age needed correcting rather than other fields, or querying large differences between the ages of spouses or partners. Nevertheless, a few households contained multiple errors which would have been difficult to resolve accurately by any automatic editing system.

Late changes to the questionnaire design had some impact on EDIS. Splitting the qualifications questions into two parts, which occurred after the 1999 Rehearsal, meant that new rules had to be devised for 2001 to deal with professional qualifications. This turned out to be the question having the largest non-response rate as many people considered that it did not apply to them.

As extra room had to be found on the form for qualifications, the question on work last week was squashed into a smaller space. As a result, two of the bullet points which appeared separately on the Rehearsal form were conflated into one, and it appears from the pattern of responses to this and later questions that some form-fillers misunderstood the question and answered No when they were actually in work. A resolution was found to this problem by amending the filter rules for the derivation of Activity Last Week, but a small number of answers may have been miscoded as a result of the extra complication which was introduced.

A single edit and imputation system was designed to deal with the censuses in England, Wales, Scotland and Northern Ireland, which all had slightly different requirements. Variations in the design of the Census form and in editing requirements meant that great attention had to be devoted to ensuring that the processing for each country was carried out to the desired standards.

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 between Rehearsal and the final version of the Census. However, these issues were identified at an early stage of Census processing.

Further results on the performance of EDIS will be reported in the 2001 Census Quality Report, which is due to be published later this year.

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