

Census 2001 Review and Evaluation

September 2003

Quality of Data Capture and Coding: 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
Background.....	2
Methodology.....	3
The ONS checks on Data Quality.....	4
Coding.....	5
Results from the LM QA System and the ONS checks of coded data.....	6
Lessons Learnt.....	17
Conclusions.....	18

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

Background

In 1997 the Census Processing team built a system to test the concept of a fully integrated and automated data capture and coding operation. The test team used OMR (Optical Mark Recognition) and OCR (Optical Character Recognition) software and manual correction keying for data capture followed by automatic coding for all text responses. Responses for which no automatic match was found were manually coded. The trial proved that this operational design was cost effective, would allow all data to be coded (compared with a 10% sample in previous censuses) and improved the consistency in output data.

It was decided that for the 2001 Census, the task of setting up the Processing Operation to extract data from the census forms and code text responses would be outsourced, and after following a negotiated route for almost a year with three prospective suppliers, the contract for processing the 2001 Census forms was let to Lockheed Martin (LM). LM decided to base their system design on the ONS model for 1997, using scanners, optical mark and optical character recognition, followed by automatic and manual coding for text responses.

The Processing Operation was to capture the data from 27.3 million forms and code the data for 59 million people. The activities at the processing site included:

- warehousing completed forms;
- transporting forms to the guillotine and scanning halls;
- guillotining the spines, and scanning the forms;
- capturing ticks and text; and
- coding text responses.

There were 18.453 billion tick boxes and 6.099 billion characters captured from the forms, and of these, 207 million tick boxes and 1.057 billion characters were sent to keyers for correction – around 1% of the tick boxes and 17% of characters captured. A significant number of these came from the high level of confidence needed for capture of form identity and date of birth – the confidence threshold for these characters was set very

high, forcing any that fell below the confidence level into keying. The keying total also includes the sample selected as part of the Quality Assurance procedures.

The total number of text responses coded was 94.68million, with 68.04million coded automatically - 71.86% over all questions. The sample selected as part of the Quality Assurance system accounted for approximately 2% of the total coded manually (frontline and expert).

Central to this huge data capture exercise was confidence in the quality of the output data. An important part of the contract with LM was a requirement that they would measure, monitor and report on the quality of data, and Service Levels Agreements (SLAs) for each type of capture and coding were set within the contract. LM produced a Data Quality Management Plan (DQMP) that detailed how they would measure and report the accuracy of their processes, identify systematic error caused during data capture or coding, and what steps they would take if accuracy fell below the agreed Service Levels.

The purpose of this report is to:

- describe the methods used by LM and ONS to measure the quality of data;
- compare the accuracy reported by LM for the delivered data, with that measured independently by ONS;
- identify successes of the capture and coding systems;
- identify particular problems in the data caused by the capture or coding processes;
- discuss the reasons for variations between the results of the LM and ONS quality assurance systems;
- discuss errors and anomalies in the delivered data; and
- recommend changes that would improve overall data quality during any future processing operation.

Census 2001 Review and Evaluation

The results quoted in this report are for the UK, as samples were taken from data for each country, including Scotland and Northern Ireland, and the results aggregated. The text relates to England and Wales only.

It must be noted that this report is an account of the history of the Quality Assurance processes and analyses. The accuracy reported for data capture was at an early stage in processing, and corrections were made during some of the following processes, so this report does not directly reflect the quality of the data that is output. Further improvements in accuracy were made and these are reflected in the published data.

Methodology

The Lockheed Martin Quality Assurance System

The Data Quality Management Plan (DQMP) described the systems and procedures LM would put in place to measure and report accuracy, the steps they would take if accuracy fell below the Service Levels, or if they identified systematic error in the data. Although the document formed part of the LM system specification, it incorporated changes requested by ONS after the Census Rehearsal.

LM incorporated an automatic Quality Assurance (QA) system within the data capture and coding subsystems. As part of the joint development and test team, ONS took the lead in testing the QA system to ensure that the counts were being taken and reported in line with the DQMP. These tests isolated faults in the system that were critical to the success of the QA and reporting processes. The tests were iterative until June 2001, when all major faults had been corrected.

A Data Quality Management Integrated Product Team was set up and held weekly teleconferences. This team comprised Data Quality and Coding managers at the Processing Site, the Contractor's System Architect and Senior Developer in Gaithersburg, and the Data Quality managers at Census HQ. Representatives from the General Register Office for Scotland (GROS) and the Northern Ireland Statistical Research Agency (NISRA) joined the meetings on an ad hoc basis. The meetings became a forum to discuss particular issues with data

quality, such as issues on the capture of form identity, the effect of black lines on images, as well as weekly progress and commentary on coding accuracy. This joint team introduced changes that improved data accuracy or operational efficiency.

The DQMP detailed the sample sizes for each data item, the formula used for calculating accuracy and the reports that would be produced. The basis for measuring and reporting accuracy relied on consistency between at least two independent sources as a proxy for accuracy. For example, if automatic coding and a manual coder selected the same code, then that was counted as accurate. If two coders selected the same code, but it was different from the automatically coded selection, then the code selected by automatic coding would be counted as an error. If automatic coding, and two manual coders each selected a different code, then this would be counted as an error. The QA process could not identify when two sources had selected the same code that was incorrect for the description given (systematic error). When an error was found, the QA process did not change the code that was output, as the number of records changed would be small, there would be no guarantee that the "new" code would be correct, and the effect on the overall accuracy of data was regarded as minimal.

Samples were taken for:

- marks and characters recognised automatically;
- marks and characters that were keyed; and
- descriptions that were coded automatically or manually by frontline and expert coders.

Data capture samples were selected for each Census District (CD) which represents a population of approximately 11,000 households. The sample size for marks in tick boxes was 0.004% which produced an average of 278 fields for QA of tick boxes that had been recognised automatically. The sample for mark boxes that were keyed, because they could not be automatically recognised, was 0.4% , anticipating an average sample of 278 per CD. The rationale for this was that of an average 6.8million tick boxes in a CD, Optical Mark Recognition (OMR) would successfully recognise 99% of tick boxes.

Census 2001 Review and Evaluation

The sample was sent to the keying staff, and operators keyed the contents of the box, which could either be empty or contain a mark, without knowing whether it was “live” data or a QA sample. If the operator keyed the same result as OMR, then OMR was credited with accuracy. If the operator disagreed, the field would be sent to a second operator. If the second operator keyed the same as OMR, then OMR was credited as a pass, but if they had the same result as the first keyer, OMR was marked as failed, and was counted in the calculation of error.

A similar process was followed for all types of data capture and for automatic and manual coding, although with different sample sizes.

The minimum sample size for fields that needed coding was 2% of records for questions on Industry, Occupation and Workplace. These are questions with high volumes of responses - almost 30 million for Occupation, 27 million for Industry and 22 million for workplace. There are around 30,000 people in a CD and Service levels were set at this level of geography, for these questions, as they produced a sufficiently high sample to assess accuracy and error rates.

The remaining questions were low volume, with Country of Birth producing 3.8 million, Ethnic group 3.9 million, Religion 1 million and Address 1 year ago 4.7 million responses for coding. Responses to these questions were infrequent at CD level, and the sample rates required to produce statistically significant data would be prohibitively expensive for processing. So the sample was 2% or 40 records, whichever was the larger and the sample was applied at Estimation Area (EA) level (around 500,000 people), and accuracy calculated and reported at that level. SLAs were not set for the different stages of capture and coding. For example the accuracy for capture of numerics was not set separately for OCR and keying, it was an overall SLA after all capture processes were complete. Similarly, there was one SLA for coding Country of Birth, regardless of whether the records had been through automatic, frontline, or expert coding.

As each EA was completed the LM system generated a variety of reports, two of which related to the quality of data capture and coding. The Data Capture report provided a list of CDs in the EA, and, by SLA, the counts for the number of characters:

- captured by OCR;
- sent for QA of OCR;
- keyed;
- sent for QA of keying; and
- totals for the EA.

The report also provided:

- a calculation of accuracy for OCR;
- a calculation of accuracy for keying; and
- the overall accuracy for the EA.

Similarly, the coding report provided corresponding counts for each question, broken down by automatic, frontline and expert coding.

The reports provided the confirmation that the sample sizes were in line with the DQMP, that accuracy was being maintained, or pinpointed CDs or EAs where accuracy had fallen below the SLA. These reports were the primary management tools for the LM and ONS data quality managers.

The ONS Checks on Data Quality

Data Capture

Following the Census Rehearsal in 1999, a thorough check of captured data was carried out. The check involved manually extracting data from paper forms or images, and comparing that with the data that was exported from the LM system. This was not just a check on the capture of the characters and marks, but of the checks, validation, and multi-ticking rules that formed part of the requirements. This was a laborious and lengthy process, but the results showed that accuracy for data capture was exceeding the service levels. This provided confidence in the quality of the data capture process, and reflected the content of the Data Capture reports produced by the LM QA system.

Census 2001 Review and Evaluation

Further rigorous testing of the system and the revised report layouts satisfied our requirements that LM were accurately measuring and reporting the counts of characters captured, the type of capture, and the accuracy achieved.

A Data Quality Management System (DQMS) was developed and run within ONS and it produced a variety of validations, counts and distributions for the tick box questions. The DQMS team could also interrogate on many other variables to identify anomalies in distribution, which could, in turn, identify systematic error in data capture.

The level of confidence gained from detailed checks of rehearsal data and the number of checks carried out by the DQMS was the basis for deciding not to carry out any manual checks on the accuracy of Data Capture during the 2001 Census.

Coding

Although there was confidence that the QA system that LM had developed would reflect the counts and accuracy as measured, there was no method by which they could identify systematic error in individual codes. This is where coders may have each selected the same code, but it was incorrect for the response given. This was particularly important in Industry and Occupation where a systematic error could affect a whole group within the classification, or in coding Ethnic Group, where an error could have excluded a whole group from the output data. It was agreed that ONS take on responsibility for checking accuracy at the individual code level, and that errors reported would be examined by the LM coding development manager, and changes made to indexes and systems where necessary.

ONS built a system to extract a random sample of coded records for each question for each EA. The accuracy of the records coded was checked by a team at ONS who were specially trained in coding each of the questions on the Census form. This team extracted samples of coded data as each EA was delivered to ONS.

The sample could be changed to meet fluctuations in the flow of work, or to allow more detailed checking if errors were found. The data presented to the coders

allowed them to reverse code the descriptions i.e look up the description of the code in the classification, and assess whether the description on the data is the same, or enter a postcode and check the address presented against that on the data. The result of their check was recorded on the system and reports produced for each EA. If particular problems were identified, further samples and analyses were carried out to pinpoint the cause of the problem and agree corrective action.

The intention was to check:

- 10% for Country of Birth, Ethnic Group, and Religion;
- all the Enumeration addresses that had the postcode allocated by the coding system (as opposed to being generated from the Form Identity); and
- 2% for Industry and Occupation, Address one Year ago and for Workplace.

The number of EAs and percentage of data checked each week was dictated by the delivery schedule, which was slow for the first five months of processing, then increased to a maximum of ten EAs in one week. The original intention was to check some data for each EA, but this was not possible with a small team of coders.

The delay in the delivery schedule provided the opportunity to carry out a complete check (100%) on all questions for the first two EAs. These thorough checks identified the majority of serious problems and, although checks were carried out on each delivery of data, throughout the processing timetable, the number of questions checked and the percentage sample varied according to the delivery schedule, staff availability and the size of the EAs delivered. The same errors were evident in the data until the first set of corrections became effective. Unfortunately some 33 EAs were being, or had completed the coding stage before corrections were implemented. The effect of this on the overall accuracy rate is shown in the tables at 3.5 and 3.6. No significant new problems were found in subsequent EAs, and the system corrections, coupled with the increasing experience of the coders, helped improve accuracy over time.

Census 2001 Review and Evaluation

The number of EAs checked for each question were:

Question	Number of EAs checked	Question	Number of EAs checked
Country of Birth	36	Enumeration Address	55
Ethnic Group	34	Workplace Address	38
Religion	30	Address 1 year ago	36
Industry	49	Occupation	44

Results from the LM QA System and the ONS checks of Coded data

Data Capture

The table shows the results for the six service levels relating to quality of data captured from the forms. The quality of capture was excellent for marks and characters, and exceeded the SLAs but the accuracy of capture for form identity as a field was slightly lower than the SLA. This, and the other issues surrounding the capture processes are discussed below:

Type of Capture	SLA %	LM Reported Accuracy %	Type of Response
OMR	99.3	99.84	Tick Boxes
OCR Alpha	96.0	99.03	Alpha characters
OCR Alpha numeric	95.0	98.99	Address and postcodes characters
OCR Numeric	98.0	99.75	Numeric characters for hours, rooms, year last worked
Date of Birth	99.5	99.93	Separate SLA @ character level
Form Identity	99.995	99.68	Separate SLA @ character level

Form Identity

The SLA for form identity was set at 99.995% which only allowed five characters in every 100,000 to be captured incorrectly. This would allow an error rate of 0.005% at the character level or 1 invalid form identity per 1,429 (a form id contained 14 characters on average).

The form identity was critical as it generated the postcode of enumeration for 94% of forms. It was vital that the enumerators' handwriting was clear and that the capture of all characters was accurate, since this was essential in placing the form in the correct output area.

In fact, there were many forms where either the enumerators' handwriting was not clear, or where clear handwriting was captured incorrectly. There were two possible outcomes to these conditions:

- An invalid form identity.
- A valid form identity - wrongly captured.

Invalid form identities were referred for correction to the Census team at the Processing office. Valid form identities that were captured incorrectly would only be identified if it created a duplicate form in an Enumeration District (ED).

For the UK, around 374,000 invalid form identities or duplicate form numbers were referred to the Processing team at Widnes – an error rate of 1.4%, far more than had been expected. Of these 13% were due to bad capture, 17% were true duplicate form numbers (a condition caused in the field that had been anticipated) and 70% due to poor handwriting by the enumerators. The effect of the errors in form identity was often to move forms into the wrong output area. The following table shows the geographical effect of the errors - an ED covered around 250 households, with an average of 46 EDs in a CD and 22 CDs in an EA.

Census 2001 Review and Evaluation

Impact	Total	Field Actions %	Data Capture %
Different form number, same ED	238,000	89	11
Different ED, same CD	78,000	75	25
Different CD, same EA	3,400	82	18
Different EA	4,500	56	44
Totals	373,900	87	13

Apart from the records referred to as invalid or duplicate, there were form identities captured incorrectly that were not invalid, and did not create duplicate records. On request, a report was developed by LM, which formed a base to investigate the incidence of this problem. The report could not be run until the forms had been through the export process and were ready for delivery to ONS, and any corrections would have delayed the delivery schedule.

This report listed forms where the identity of the box containing the form was not consistent with the form id. Since the box identity and part of the form identity should be the same, the report listed forms that were either in the wrong box, or the identity had been captured incorrectly. The majority of form identities listed on the report had been captured correctly, but many of the box identities had been poorly written, with subsequent poor capture, and these were excluded from further analysis. ONS analysed the smaller number of forms that seemed to be in the wrong box, and these were forms where the identity captured was valid, but was not actually the identity written on the form. Data was delivered by EA, and prior to delivery any form identities that were duplicated would be listed, and the form identity and postcode corrected before the data was delivered. There were some forms where the postcode allocated was inconsistent with the area of enumeration, and these were investigated and corrected by ONS geography after the data was delivered. The

incidence and impact of the remaining invalid form identities did not warrant correction and the subsequent delay to delivery.

Poor handwriting and poor capture contributed to the failure of LM to meet the service level in 80% of CDs. The overall average accuracy reported was 99.68%, an error rate of 0.32%. This is different from the error rate of around 1.4% that comes from analysing the duplicate and invalid form identities, and highlights the need to find better ways to capture, and measure the quality of, form identity. The changes made after the data was captured improved the overall quality for this field, although we have no actual measure of output accuracy.

Date of Birth

Age is a critical field for Census, as many output tables are based on sex and age. Age is derived from Date of Birth (DoB), and so it assumes the same importance. The SLA for DoB processing was 99.5% at the character level. There were 469.77 million characters captured for DoB, with a potential of 2.34 million allowable errors in the data.

LM assessed overall accuracy as 99.93%. ONS assessment of accuracy at character level was 99.8% (above the SLA), producing 406,000 errors in DoB. The impact of an error depends on its character position; an error in year of birth has a bigger impact on the calculation of age than an error in the day or month. Errors in year of birth can turn a child into a pensioner for example. These errors would subsequently have values imputed at the edit and imputation stage and lower the overall quality of the data at output.

In hindsight there are various steps that could be taken to improve the quality for this field:

- the SLA could have been set at Field level;
- we could have had a service level purely for capture of year of birth, and leave day and month as regular numeric fields for QA; or
- introduce a consistency check during capture and correct year of birth before imputation.

Census 2001 Review and Evaluation

Part of the Enumeration District (ED) audit was to check each record in selected EDs to see whether the captured information was the same as on the form image for DoB. The results of this check were analysed for 64 EDs and used to assess accuracy:

- 28,561 person records checked.
- 228,448 characters checked.
- 0.14% characters captured incorrectly.
- 0.69% fields captured incorrectly – field accuracy 99.31%.
- 0.08% of errors caused by a system error where rules were not applied consistently during the capture stage.

These sample percentages suggest around 406,000 records in total with errors somewhere in date of birth. The impact of the errors can be categorised into high, medium and low:

- High Impact - 100,000 records - 47,000 records where the rules governing capture of the unit number in year of birth were applied incorrectly. 53,000 records where the error was made in capturing the decade in year of birth. This, for example, could change a schoolchild into an older person, resulting in imputed values for family composition, industry, occupation and qualifications.
- Medium Impact - 29,000 records with errors at the unit level in year of birth – this becomes important if the age is tripped into a different 5 year age group.
- Low Impact - 270,600 records with errors in day or month. These fields only have an impact if the day of birth is around Census Day in April, as the age can differ by one year if the day of birth falls before or after Census day.

Black Lines on Images

As data was delivered and integrity checks carried out, it became clear that black lines were appearing on images when no data existed. When they appeared, the lines crossed through tick boxes and additional person records were created. The lines also crossed through boxes that were for a real person and had been left blank on purpose. These spurious lines produced values that were inconsistent with the rest of the data for that record.

There were two main causes of the problem:

- dust from forms gathering on the scanner being 'seen' as a mark on the form; and
- form fillers crossing through all the pages they didn't complete.

The errors did not affect the overall SLA, which was set at the CD level for tick boxes, but they did affect statistics for some smaller areas. An edit was carried out as data was delivered to identify spurious records and data that had been created. For a record to be regarded as a true person, two of four data items had to be completed on the form – the four items were Name, Date of Birth, Marital Status and Sex. The data items present had to be either Name or Date of Birth with one other of the three remaining items. A further edit was carried out to correct inconsistencies in records created by the black lines.

The majority of errors were identified and corrected automatically by these edits.

Number of Rooms

Data Capture produced an excessively high number of household records that had more than ten rooms. During the capture process, images were presented to keyers when a number in the first box was crossed through, then written again in the second box. The OCR referred the crossed through box for a keyer to validate the value, and they often keyed the numeric value, rather than a space. This resulted in the disproportionate number of households with more than ten rooms. Another edit was added to the program that identified and corrected errors before downstream processing began.

Coding

All text responses were submitted to an automatic coding system, and if not successfully coded were referred to manual coders using a Computer Assisted Coding system. As each EA completed coding, LM produced a report giving counts of records coded, sample sizes and accuracy for automatic, frontline, and expert coding.

Census 2001 Review and Evaluation

The following table shows the LM report on accuracy that covered the whole of the processing schedule, together with the ONS assessment of accuracy based on the results of checks carried out, after delivery, on samples of coded data. For all but two address questions (Enumeration and Workplace) the accuracy assessed by ONS confirmed that LM had met or exceeded the service levels, confirming that consistency is a good measure of accuracy. The differences between the LM and ONS results are discussed below, as are some particular problems reported for each question.

The volume of text responses for Country of Birth, Ethnic Group and Religion are relatively small, as the majority of respondents would have ticked a box. There were no tick box options for Industry and Occupation, and this accounted for the high volume of coding for these questions.

Question	Number Coded	% of Population	SLA %	ONS Assessment Accuracy %	LM Reported Accuracy %	LM Automatic Coded %
Country of Birth	3,780,151	6.36	96.0	99.80	99.8	81.6
Ethnic Group	3,866,964	6.51	96.0	96.80	98.6	75.7
Religion	1,045,874	1.76	96.0	97.00	98.8	74.4
Industry	27,970,005	47.08	88.0	88.22	89.1	66.8
Occupation	29,611,276	49.85	88.0	89.04	91.1	72.2
Enumeration	1,651,445	2.78 (<i>h/h</i>)	100.0	98.51	99.3	88.4
Workplace	22,056,446	37.13	94.5	86.12	94.3	71.8
Address 1 year Ago	4,720,878	7.95	96.5	96.19	98.1	83.6

Except for Occupation, responses were coded independently from other questions, so generally there were no consistency checks between questions. For example there was no check to see whether an “Officer” in the army had an industry of “Defence Activities”. However there are some occupation titles for which the code is different dependant on the industry or qualifications given – for example a doctor must have a certain level of qualification, and the coding system was developed to handle those cases.

Industry

Historically the information provided by respondents for “business of employer” (Industry) has not provided sufficient detail to code at the lowest level in the classification. LM based their estimates for throughput

and manual coding staff on attaining a certain match rate for automatic coding, and during development of the system they realised that they were not going to achieve that rate, although they were exceeding the SLA for accuracy. In order to improve the automatic match rate, LM made use of the SIC (Standard Industrial Classification) code where it was available on the business address file used to code workplace address. This increased throughput for automatic coding by around 10%. The drawback to using this source was to reduce the overall average accuracy (see table below), and until LM were able to make adjustments to the system, the overall average accuracy fell below the SLA.

Census 2001 Review and Evaluation

	Accuracy after 24 EAs checked %	Accuracy after 31 EAs Checked %	Accuracy at close: 49 EAs Checked %	Overall Improvement %
Industry coded from text	88.53	90.01	90.95	2.42
Industry coded from Business address file	53.32	58.53	61.11	7.79
Overall accuracy	85.44	87.13	88.22	2.78

Accuracy over Time

During the course of processing ONS identified 120 individual codes exceeding the criteria that identified possible systematic error. Of these, 79 were applied automatically, and 31 applied incorrectly by manual coding.

The LM Development Manager carried out an initial investigation into the errors, and in discussion with ONS and the Coding managers, corrections were agreed. Changes were made to indexes, tuning data (real descriptions from previous census data, with correct codes) parsing rules, and coding instructions, as appropriate. The first set of changes were made when around 33 EAs had completed coding, and a further 20 changes were incorporated into the automatic system or additional guidance for coders.

LM also tuned the system to use the SIC data from the Business address file, and accuracy for this coding source improved by 7.79%, and the overall average accuracy of data improved by 2.78%.

ONS carried out a detailed check of the records coded from the Business address file that were marked as errors. This was to ensure that the errors were not concentrated in particular parts of the classification, causing skew or bias in the output. There was no evidence of bias or skew in the data.

Classification

There were 27.97 million records coded into one of 31 major categories. The SIC has Divisional, Group and then Class level of detail but the majority of responses did not provide sufficient information to code below

Group level, so the Census used a collapsed version of the classification for coding. Codes were allocated to a minimum of two characters (Division), some sections were coded to three characters (Group), and only a few to the fourth character (Class). The collapsed classification can be found on the National Statistics web site (>http://www.statistics.gov.uk/census2001/pdfs/class_section5.pdf).

Results

ONS checked a total of 364,330 industry codes - 1.3% of all records coded. Errors were logged at each character position within the code, and for industry overall. 64% of errors were in the wrong Division and 36% were in the correct Division, but with incorrect Group or Class level coding. LM reported an overall accuracy of 89.10% and ONS estimates 88.22% (a difference of 0.88%). It is likely that most of this difference is accounted for by errors identified in the first delivery of data for which no corrections to indexes etc were made until much of the data had been coded. The tables below show the results for each of the thirty one sections used in the ONS coding frame. The first shows those where ONS have measured accuracy above the SLA (88%) and the second shows the sections where LM did not achieve the SLA. Each table shows the percentage of the total sample that fell into each section. The representative percentages together produce a discrepancy of 0.23% that is due to rounding.

Census 2001 Review and Evaluation

Classes above SLA

The classes that exceeded the SLA represent 75.65% of the population coded. They are shown in the order they appear in the classification, together with the related codes.

Code	Description	ONS Assessment of Accuracy %	% of Sample Represented
10,11,12	Mining/Quarrying (energy products)	90.86	0.27
17,18	Manufacture, textiles and products	89.75	0.88
190,191,192,193	Manufacture, leather and products	88.32	0.03
210-212,220-223	Manufacture, pulp/paper/products/printing/ publishing	90.32	1.48
23	Manufacture, coke/petroleum products/nuclear fuel	90.10	0.10
340-343,350-355	Manufacturer, transport equipment	88.89	1.41
400-403,41	Electricity/Gas/Water	92.56	0.70
45	Construction	94.48	5.99
500- 505, 51, 52, 530	Wholesale/Retail	93.02	15.72
550-555	Hotels/Restaurants	96.88	5.28
600-603, 61,62, 630-634, 640-642	Transport/Storage/Communication	91.68	6.31
65-67	Financial Mediation	95.02	4.11
70-73,7411-7415, 740-748	Real Estate/Renting/ Business activities	89.91	10.64
800-801, 8021, 8022, 8031,8032,8041,8042	Education	90.83	7.63
850,8511-8514,852,853	Health and Social Work	93.81	10.78
90,91, 920-927, 93	Other Community/ Social	90.19	4.29
99	Extra Territorial organisations and bodies	90.91	0.03

Census 2001 Review and Evaluation

Classes below SLA

These sections account for less than 25% of the coded sample. Records marked as uncodeable that could have been coded represent 0.52% of the sample.

Code	Description	ONS Assessment of Accuracy %	% of Sample Represented
Various	Coded from Business address file	61.11	9.15
010-105, 02	Agriculture, hunting	85.31	1.17
05	Fishing	72.73	0.05
13,14	Mining (other)	84.21	0.08
150-159, 16	Manufacture, food/ tobacco	80.66	1.54
20	Manufacture, wood and products	87.39	0.15
240-247	Manufacture, chemicals/products/ man made fibres	85.88	0.90
250-252	Manufacture, rubber/ plastics	82.40	0.55
260-268	Manufacture, non- metallic products	87.37	0.32
27,28	Manufacture, base metals	87.92	1.45
290-297	Manufacture, machinery not elsewhere classified	82.16	0.87
30-32,330-335	Manufacture, electrical and optical	82.26	1.44
360-366,37	Manufacture, (not elsewhere classified)	68.71	1.17
750, 7511-14, 7521- 25, 753	Public Admin/defence/ social security	87.06	5.17
95	Employees in Private households	86.04	0.08

Census 2001 Review and Evaluation

Within the data there were specific problems that caused anomalies in the output:

Armed Forces

There were two main problems with code 7522 (UK defence activities) where British armed forces and civilians working for the Ministry of Defence (MoD) would normally be coded.

- A substantial number of respondents who work in electronics, building or service industries for the MoD stated “Defence Activities” as the business of their employer. Although it was clear from the occupation titles and the employers’ names that these people were not in the defence industry, they were coded to 7522 because:
 - the description matched an entry in the coding index;
 - industry was coded independently from occupation and employer’s name; and
 - there were no consistency checks between industry and occupation.
- Automatic and manual coding wrongly assumed the 7522 code for members of foreign armed forces stationed in England. All of the areas affected by this fault had passed through coding before the first affected EA was delivered. This issue was resolved during downstream processing, when a program was run to identify those in code 7522, with Country of Birth as North America. This corrected the classification for the majority of the foreign armed forces who were in England at Census time.

Knitwear and Knitted Garments

Division 17 gives manufacture of knitted pullovers, cardigans etc and Division 18 is Manufacture of Wearing apparel. The type of garment manufactured dictated the code, and as the type of garment was rarely provided, coders should have marked the records as uncodeable. If this had happened, a code would have been imputed and there would be a visible gap in the output in specific geographic areas. The decision was taken to let the coders continue allocating 17 or 18 arbitrarily, as this at least placed the respondents in the correct part of the classification, and was better than imputing values, and perhaps having gaps in the output.

Wholesale v Retail

The emergence of ‘Out of Town’ shopping centres with superstores has clouded the ability of form fillers to distinguish between wholesale and retail, and some stores sell to trade and retail customers. Form fillers gave ‘wholesale’ as business of employer, and ‘selling to public’ as the occupation description. Coding was carried out according to instructions, but we estimate that around 5% of records that were coded to wholesale really belong with retail.

Records marked Uncodeable

There were descriptions where the form filler did not supply sufficient information to allow coders to select the right code and these were, correctly, marked as ‘uncodeable’. Analysis of the uncoded records showed that some industries would not be adequately represented in the output data. e.g. ‘pharmaceuticals’ was given as the employers business, and around 71% of responses within the pharmaceutical industry had no additional information that coders could use to determine whether they were in manufacture, retail or wholesale. If no action was taken, the output data would have contained only 29% of the total records for this group.

There were a number of industries similarly affected, and to counter this, default codes were added to the manual coding instructions for each of the industries concerned. Approximately 45% of EAs had been coded before these additional steps were taken, and this may account for any shortfall in the expected numbers for the affected groups.

Occupation

Occupation is generally considered to be easier to code than Industry, even though accuracy of coding can be dependent on the industry code and the level of qualification attained. There were 29.61 million occupation titles coded into one of nine Major Groups. Coding was always to four characters.

The quality of responses for Job title was much better than Industry, allowing 72.2% to be coded automatically with the remaining 27.8% coded manually. Average manual coding speeds were 200 responses per hour – much higher than previously achieved in tests.

Census 2001 Review and Evaluation

ONS checked the codes for 355,243 responses – 1.2% of the total coded, and of these, 38,948 were incorrect. ONS logged errors at each character position for each major group and for Occupation overall - 63% of errors were in the wrong major group, with 37% in the correct Major Group, but the wrong character position.

There were 5,181 records marked as uncodeable, with 806 of these (0.022% of the total records coded) that could have been coded.

Accuracy over time

There were 12 sets of changes to the system during the processing operation. The changes were identified from investigating reported errors in 133 individual codes. The system changes incorporated additions, deletions and amendments to the indexes, tuning data, parsing rules, and coding instructions. During the processing operation the overall average accuracy of data improved by 2.2% as changes became effective.

LM reported overall accuracy as 91.1%, and ONS report 89.04% - a variance of 2.06%, this difference is due to the errors identified in the first delivery that were not corrected until much of the data had been coded.

	Accuracy after 13 EAs checked %	Accuracy after 25 EAs Checked %	Accuracy after 45 EAs Checked %	Overall Improvement %
ONS assessment	86.84	88.30	89.04	2.20

Results

The table below shows each of the nine groups, with the ONS assessment of accuracy, together with the percentage of population represented in each group.

Group Code	Group Description	ONS Assessment of Accuracy %	% of Sample Represented
1	Managers and Senior Officials	*81.57	13.20
2	Professional Occupations	*90.60	10.14
3	Associate professional/ technical	*88.19	12.82
4	Administrative/ secretarial	87.84	13.44
5	Skilled Trades	92.68	11.46
6	Personal Service	94.40	7.42
7	Sales and Customer Service	91.45	8.88
8	Process, Plant, Machine operatives	*84.70	8.88
9	Elementary Occupations	92.16	13.76

*Groups affected by the systematic error for qualified occupations (see below).

Qualified Occupations

In some cases the coding system allocated codes for professions requiring educational qualifications to respondents who had no qualifications, and codes for occupations that do not incorporate qualifications to respondents with professional qualifications. There were four groups affected by this error - Managers, Professional Occupations, Associate professional/ technical, and Process Plant and Machine Operatives. A program fix was developed and applied to the system, by which time 33 EAs had either been completed or were in the process of being coded. The skew in the data caused by this error was considered severe, and ONS identified and corrected 7,500 records affected in the EAs processed before the program was fixed.

A further complete check on all the records that could have been affected by this error was carried on the first data coded after the system fix was made. This check confirmed that the systematic error had been corrected.

Census 2001 Review and Evaluation

The overall accuracy shown in the table above for the affected groups reflects the systematic error in the delivered data, with Managers at 81.57%, and Process, Plant and Machine Operatives at 84.7%. This accuracy rate does not reflect the corrections made to the data after delivery.

Armed Forces

The Ministry of Defence (MoD) were consulted about the way the armed forces should be expected to describe their job title. MoD said they would instruct their staff to enter 'Officer' or 'other Rank'.

Many of the armed forces did not follow this instruction, they wrote in a standard job title such as 'radio operator', 'mechanic', 'chef'. These standard job titles were coded correctly according to the coding rules and resulted in a 20% drop in the expected count for armed forces personnel.

The job titles given incorrectly by armed forces staff were generally coded correctly by the system, and were not therefore marked as errors in the calculation of accuracy.

Although the coding team had advised downstream processing that there was a problem with armed forces personnel, the impact was not quantified until the Quality Assurance of the One Number Census (ONC) process, when expected counts of Armed forces were not confirmed in the Occupation data. A separate exercise was carried out to identify the likely armed forces personnel by combining several coded items and text responses in records where the industry code was 7522, Defence Activities, and these records are identified with a marker on the census database.

Use of 'Technician'

The word 'Technician' was over-used in job titles for the type of job being done – eg. keg technician was sometimes used for a barman/barmaid, vision technician for a window cleaner, floor technician for a cleaner. In the Index, the Technician group are associate professional and technical jobs, whereas the descriptions given by the examples quoted showed that they were actually barmen, window cleaners, and cleaners. The

titles given were coded according to instructions, and led to all the above examples being coded to a higher level in the classification than the one to which they really belonged.

Volume of Uncodeable

There were descriptions where the form filler did not supply sufficient information to allow coders to select the right code and these were, correctly, marked as 'uncodeable'. Analysis of the uncoded records showed that some occupations would not be adequately represented in the output data. e.g. there are over forty different codes for Builders, and without supporting details from the industry or occupation description, there was no way of knowing which code to select. In the first EA checked there were 130 "builders", 43 were coded and 87 were marked as "uncodeable". For this, and other groups that were similarly affected, ONS added default codes to the manual coding instructions. Approximately 45% of EAs had been coded before these additional steps were taken, and this may account for any perceived shortfall in the output counts.

Enumeration Address

Postcode of enumeration was the key for allocating forms to census output areas and, for forms where the postcode was not derived from the form identity, the postcode was either captured from the form, or coded from the address provided. Approximately 94% of forms had their postcode derived from form identity, leaving around 1.65 million to be coded by other methods. Of this remainder, 1.1 million were accepted by checking that the postcode on the form was valid and correct for the ED.

The addresses for the remainder, approximately 0.55 million, were captured and passed into the automatic address coding system where they were subject to the QA process. The sample selected was 2% for automatic coding and 100% for manual coding.

The ONS assessment of accuracy is therefore based on the 0.55 million records that entered the LM coding system. All other postcodes were regarded as accurate. The error rate for coded records was 1.5% , which suggests around 6,900 (0.026%) records with errors in enumeration postcode.

Census 2001 Review and Evaluation

ONS accepted valid postcodes that were captured from the form. LM included the counts of these postcodes on the Coding Report and attributed them with 100% accuracy within their calculation. The accuracy measured by ONS applies to the records that were actually coded by LM, and this different approach accounts for the variation in results.

ONS had frozen the address lists about 3 years before Census day, so new addresses after that date did not have a unique identifier, and therefore increased the volume of records sent to the coding system. The combined effect of this, with LM being unable to link postcode to Ward code during manual coding, obliged us to accept a large number of postcodes that were either not in the same area as the form, or were not on the pre-listed address file. ONS Census Geography checked each of these records, correcting the postcode where appropriate.

Workplace Address

There were just over 22 million records with responses to workplace address. LM coded 13.5 million and the remaining 8.5 million (38.6%) had valid postcodes that were captured from the form.

The contract with LM allowed them to provide partial postcodes for this question, and the samples of data checked included full and partial postcodes and records marked as uncodeable.

The ONS assessment of accuracy was 86.12%, compared with the LM report of 94.30%. This is the greatest variation between what LM reported and ONS regarded as accurate. If the assessments for partial postcodes and uncodeables were discounted, then the overall accuracy rose by 3% to 89%. The remaining discrepancy of 5.3% is due to the 38.7% of records where a valid postcode was captured from the form, and counted as 100% accurate in the LM coding report. ONS analysed the errors and the majority of incorrect postcodes were generally in the correct geographical area, but wrong in the last two character positions, and for workplace statistics this would be sufficiently accurate.

The level of accuracy needed for this question, and the difficulty in coding it should be taken into account when setting the Service Levels for any future contract.

Migration (Address 1 Year Ago)

The high accuracy of 96.19% reflects public awareness of residential addresses, and the better coverage of residential addresses in postcode software. There were 4.72 million records for address one year ago, for which respondents provided postcodes for 2.97 million (62.92%) were captured from the forms.

Ethnic Group

From a coding perspective, Ethnic Group is a low volume question, as most people respond via the tick boxes. However, it has a high public profile. At 3.87 million (6.51% of population) the number of written responses was far higher than the estimated 3%. In Wales the number was 15.7% where people responded as Welsh, which had no pre-defined tick box.

The main difficulty with Ethnic Group coding was making a determination on mixed group, and in coding Asian and mixed Asian groups. This was a consistent error with manual coders. ONS advised LM that there were problems in the following codes:

Code Allocated and Description	Correct Code and Description
59 other Asian, not specified	41 - Indian or British Indian,
59 other Asian, not specified	42 - Pakistani, British Pakistani
29 mix unspecified	37- European mixed
29 mix unspecified	90 - no mix
90 no mix	29 - mixed, not specified
various mixed	29 - mixed, not specified

These errors persisted throughout processing - mostly affecting the Asian groups. The overall effect is a shortfall in the specific groups of Indian, Pakistani and Bangladeshi, and a large number in code 59 - Other Asian and Asian unspecified.

Census 2001 Review and Evaluation

Country of Birth

This question had a low volume of written responses – again, most respondents ticked a box. It had been estimated that around 4.7million (8% of population) would provide a written response for coding, but the number only reached 3.78 million (6.36%). The auto coding system allocated codes using a direct match process for 81.6% of responses, leaving a small number for manual coding. Most of those requiring manual coding were due to spelling mistakes. There were no particular recurring errors to report. Our assessment of overall accuracy was 99.8%, the same as LM assessment, and 3.8% above SLA.

Religion

This was the first time Religion was asked in a full census, so there was little information on which to base the estimate of 900,000 responses. Again, most respondents ticked a box and the actual number of text responses was 1.045 million. There were no particular problems in coding the responses, although following the release of a new Star Wars film, many respondents claimed Jedi as their religion, and a surprising number who entered TOG (Terry's Old Geezer – a follower of a popular Irish radio and TV presenter). Jedi was allocated to a unique code, so that the volume could be monitored. TOGs were included in the “other” category.

Lessons Learnt

- The quality of captured and coded data met or exceeded the quality requirements except for two critical fields - Form identity and Postcode of Enumeration address. There is no true measure of accuracy for either of these two fields that are crucial to placing households in the correct output area. Any future contract should incorporate quality standards needed for all stages of processing, and these standards must be measurable and achievable. There must also be an opportunity to correct errors at source and to reflect the effect of the corrections in the reported accuracy.

- ONS must remove the reliance on enumerator handwriting to provide a form's unique identity – and therefore its link to an output area.
- Operational processes affected the quality of the data. Guillotines and scanners should not be located in the same area. Action to be taken by operators if they witness black lines during their quality check, and cleaning schedules for scanners should be included in any future contract. These actions should be audited as part of the processing team responsibilities.
- The QA processes developed by LM and the ONS assurance processes worked well, complemented each other and delivered data that improved in accuracy during the processing schedule. All major problems in coded data were identified in a comprehensive check of the first EA. Any future contract should incorporate a requirement to have a large block of data delivered early enough for checks to be carried out and corrections to be applied before much more of the data is coded.
- The contractor should have had systems in place to identify errors at the individual code level, and then analyse the impact of these in the data. This should be a mandatory requirement in any future contract.
- The level of detail needed for Coding should be reviewed in the light of the poor quality of response for industry, and more default codes should be considered for groups that are recognised as particularly difficult to code.
- Set SLAs at field level for critical fields.

Census 2001 Review and Evaluation

Conclusions

- The quality of the captured and coded data was excellent. This is the first time that QA processes have been measurable and have continued as an automated background activity for the whole of the processing operation. Confidence in the system stemmed from a robust testing strategy developed within ONS, and carried out by ONS with a high level of commitment of system and staff resources from the contractor.
- Consistency is a good measure for accuracy except where descriptions are coded consistently, but inaccurately. Checking a sample of coded data was an effective method of identifying these systematic errors, assessing reported accuracy and improving accuracy over time.
- The LM System Architect, Coding Development manager, Data Quality manager, and ONS processing staff provided continuity, experience and expertise throughout the development and operational stages. This ensured an ongoing commitment to delivering a successful QA system and to improving the quality of captured and coded data.

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