2001 CENSUS: INDUSTRY

(Census Paper No. 03/08)

2001 CENSUS: INDUSTRY

(Census Paper No. 03/08)

Liz Costello

Population Census Evaluation
February 2004

INQUIRIES

For further information about this paper, contact the Assistant Director, Census Evaluation by telephone: (02) 6252 5611 or email: <joanne.healey@abs.gov.au>.

SUMMARY OF FINDINGS

This Industry Census Paper evaluates the data quality of the Industry questions in the 2001 Census. Topics analysed include: changes made to the Industry questions and the coding procedures between the 1996 and the 2001 Censuses; non-response rates; levels of undefined coding and coding discrepancies; a comparison with the August 2001 Labour Force Survey; and possible changes for the 2006 Census. The main conclusions of the analyses are:

- The non-response rate for Industry of employment in 2001 was 1.7 per cent, a slight improvement on 2.0 per cent recorded in 1996. When compared to other labour force-related variables, Industry had the third highest response rate after Occupation, which had a 1.2 per cent non-response rate and Job Last Week, a 1.4 per cent non-response rate. The non-response rate for Industry increases with age, which is consistent with the response rates for other labour force-related variables.

- An average of 55.1 per cent of responses were coded by the Automatic Coding (AC) system, leaving 44.9 per cent processed by Computer Assisted Coding (CAC) and Query Resolution (QR) processes. The Industry division, Education had the highest AC rate, with 77.2 per cent and the Industry division, Manufacturing had the lowest, with 43.7 per cent.

- There were 8,298,606 applicable Industry responses of which 1,355,093 (16.3 per cent) were subject to Quality Management (QM) coding. Altogether, 70,465 Industry discrepancies (5.2 per cent) were recorded in the Management Information System (MIS) reports.

- The Industry division Transport and Storage contained the highest level of undefined coding with 79.5 per cent of the responses coded to the ANZSIC class level. Manufacturing division recorded the second highest level with 83.0 per cent of responses coded to the most defined level. The lowest levels of undefined coding occurred in the Government Administration and Defence and Personal and Other Services both with 99.5 per cent of the responses coded to the ANZSIC class level. The most significant improvements between the 1996 and 2001 Censuses in responses coded to the most detailed ANZSIC level occurred in Agriculture, Forestry and Fishing (up 26.1 percentage points) and Mining (up 21.1 percentage points).

- The data reconciliation between the 2001 Census and the August 2001 Labour Force Survey showed that the differences in estimates between the two collections were statistically significant as was the outcome of the comparison of the two collections for 1996.

- For the 2006 Census, the ABS will be looking at rewording the Industry question to more closely align Industry responses with ANZSIC classification principles.

- For 2006, industry responses will be dual coded, in the first instance using the 1993 ANZSIC, and secondly on the basis of the new 2006 ANZSIC which is currently under development.

CONTENTS

LIST OF TABLES AND FIGURES

1. INTRODUCTION

1.1 About Census Papers

The Australian Bureau of Statistics (ABS) has a stated, corporate objective to provide the means for informed and increased use of statistics. This Paper is one of a series produced after each Census by the ABS Population Census Evaluation team, whose role is to review the data quality of the 5-yearly Census of Population and Housing.

Census Papers aim to inform users of issues identified as impacting on the quality of the census data, that they should keep in mind when utilising the data. Analyses such as these are a critical factor in the continuous quality improvement of the Census Program.

The ABS welcomes your feedback and suggestions.

1.1.1 This Paper

The focus of this Paper is Industry of employment data which have been collected in all Australian Censuses since 1911.

This Paper discusses the quality of Industry data collected in the 2001 Census and contains:

- a description of Industry coding procedures used in the 2001 Census and data quality issues associated with those procedures;
- an analysis of the impact of Intelligent Character Recognition (ICR) technology on Industry data;
- an analysis of the frequency of undefined coding of Industry data in the 1996 and 2001 Censuses;
- an analysis of non-response rates for Industry data;
- an analysis of Industry coding discrepancies;
- a data comparison between 2001 Census and August 2001 Labour Force Survey Industry data; and
- changes being tested for the 2006 Census.

The monthly ABS Labour Force Survey in which employed persons are asked for their Industry of employment each quarter, is used for comparison in this Paper. Industry data at the ANZSIC group level is available from the survey but some data are subject to quite high sampling variability. Demographic characteristics of the employed and unemployed are also collected each month. Industry data from the 2001 Census is compared to Industry data from the August 2001 Labour Force Survey.

For intercensal analysis, 1996 data has been obtained from an equivalent population. Therefore, some of the figures quoted in this paper may differ from those in the paper titled, *1996 Census Data Quality: Industry (Working Paper No. 00/3).*

1.2 Background

Industry was initially coded on the basis of the response given to an Industry description question. From the 1954 Census until the 1996 Census, in addition to an Industry description

question, a question has asked for the employer's name and address. From the 1971 Census through to the 1996 Census, employer's name and address responses were used as the first attempt to allocate an Industry code by matching this information to businesses listed on a subset of the ABS Business Register, a comprehensive list of Australian businesses coded by Industry classification. This process was known as business matching. Information from the Industry description question was used only where it was not possible to match the employer's details to an entry on the Business Register, a process referred to as Industry description coding.

Soon after the 1996 Census, it was decided to adopt a Structured Coding Methodology for Industry coding because of concerns about the availability of business information at location level in 2001 and the desire to make the coding of Industry responses more consistent with the approaches used to code occupation and qualifications responses. For the 2001 Census, Industry responses were coded by the ABS Coder using the newly developed 'structured' Industry coding index. See *Section 4.4 The Industry Classification and Indexes*, for further information.

Employer address also changed to a workplace address in 2001. For more details refer to *Section 2.1 Changes to the Industry Question Format*.

Before the 1971 Census, the ABS used an internally developed Industry classification known as the Classification of Industries. From 1971 through to 1991, Industry was coded using the Australian Standard Industrial Classification (ASIC). For the 1996 Census, Industry was coded both to ASIC and to the Australian and New Zealand Standard Industrial Classification (ANZSIC) which was developed in 1993. However, 1996 Census output products relating to Industry were only available by ANZSIC. Responses to the Industry questions for the 2001 Census were classified using ANZSIC. See the *2001 Census Dictionary* (cat. no. 2901.0), for further information.

Like all other reported information in 2001, Industry employer names, and workplace addresses were destroyed once computer processing had been completed, unless the person had agreed to having their name-identified information retained for 99 years and then released in 2100 for research purposes, as part of the commemoration of the centenary of the Federation of Australia activities.

1.2.1 Purpose of the Industry Question (User Requirements)

Employment data by Industry are needed for analysing and monitoring the rate of structural change at a national and local area level. Detailed analyses are undertaken on the demographic and labour force characteristics of employees in industries and locations which are facing extensive structural change. Data on the geographic distribution of Industry of employment is needed to monitor these changes in order to provide a basis for social and economic policy and planning.

Small area and regional data about the structure of the labour market are required for the purpose of advising all levels of government, and their agencies responsible for delivering programs and providing services at a regional level.

Industry data are widely used in the analysis of the labour market. The utility of the data is considerably enhanced when analysed with detailed data on occupations and qualifications. Although a substantial amount of information on employment by Industry is available from other ABS collections, it is not available at a detailed level for most industries, for small areas or cross-classified with other employee characteristics, as is the case with Census data.

Industry Sector data coded from business names, indicate whether employment establishments are owned by the private sector or by one of the various levels of government. These data are used to assess the impact of government activity in small areas and to identify Indigenous people employed in the Community Development Employment Program (CDEP). Note: Industry Sector data is not a subject of this paper.

Names and addresses of a person's workplace are also used for the coding of work destination zones used in journey to work studies. The employer's address is used to find out what journeys people make to get to their workplace.

1.3 Changes Between the 1996 and 2001 Censuses

1.3.1 Intelligent Character Recognition (ICR)

One of the most significant changes for 2001 was the design of the Census forms to utilise ICR processing. ICR processing, along with Optical Character Recognition (OCR), scans the forms and converts mark-box, numeric and alphabetic hand-written responses to codes and text.

ICR is cost-effective technology, improving processing timeliness while delivering a high standard of data quality. An ICR approach minimises human error while maximising coding consistency and enables hand-written text and figures to be automatically deciphered and coded. ICR technology featured in all four Industry-related questions.

Details of the impact of ICR technology are discussed in *Section 4 Processing at the Data Processing Centre (DPC).*

1.3.2 Industry Question Format

A two-question design was introduced in 2001 in order to determine the person’s Industry of employment. The first question asked for a description of the employer's business, while the second asked for the main goods produced, or main services provided, by the employer’s business. The two-part Industry question was expected to improve the quality of responses by identifying the activity and products of the employer’s business rather than the broader nature of the business. Refer to *Section 2 Question Design* for more information about the 2001 Census Industry questions.

2. QUESTION DESIGN

2.1 Changes to the Industry Question Format

Both the question asking for the employer's business name and the instructions on the form relating to that question, remained unchanged from 1996 to 2001. The only change made on the 2001 Census form was to the response area which was redesigned to facilitate ICR technology. Refer to Figures 1 and 2.

FIGURE 1: EMPLOYER'S BUSINESS NAME QUESTION, 1996 CENSUS HOUSEHOLD FORM

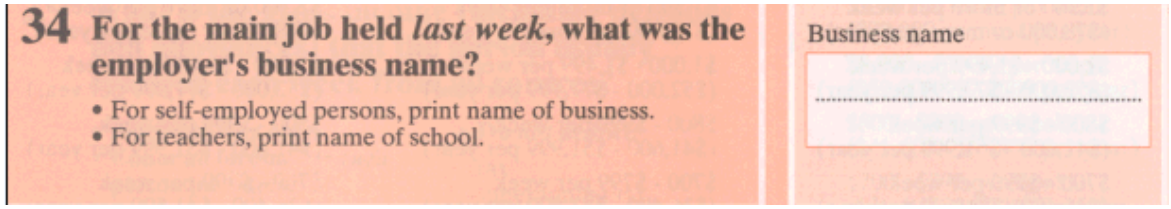
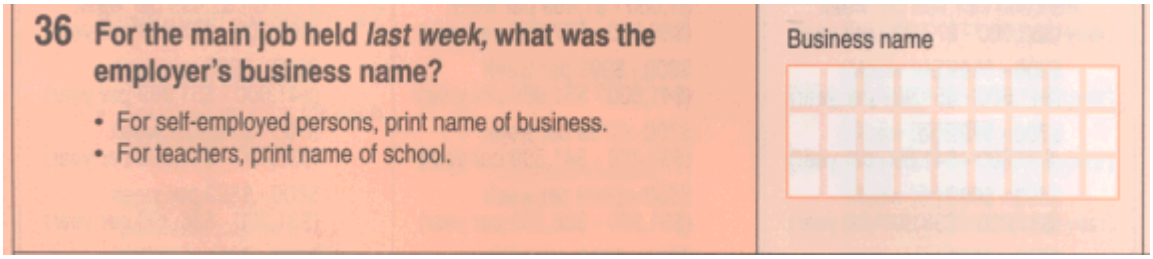


FIGURE 2: EMPLOYER'S BUSINESS NAME QUESTION, 2001 CENSUS HOUSEHOLD FORM



Changes were made to the wording of the question asking for the person's workplace address and the instructions relating to it. In 1996 the question asked for the employer's workplace address. The change in wording was intended to emphasise the person's workplace address and reduce the volume of responses that gave the employer's head-office address. In 1996 many respondents gave their employer's address, stating the head-office address rather than their actual place of work. For example, people working in schools responded with Department of Education in Sydney. The format of the response area on the form was also redesigned. Refer to Figures 3 and 4.

FIGURE 3: EMPLOYER'S WORKPLACE ADDRESS QUESTION, 1996 CENSUS HOUSEHOLD FORM

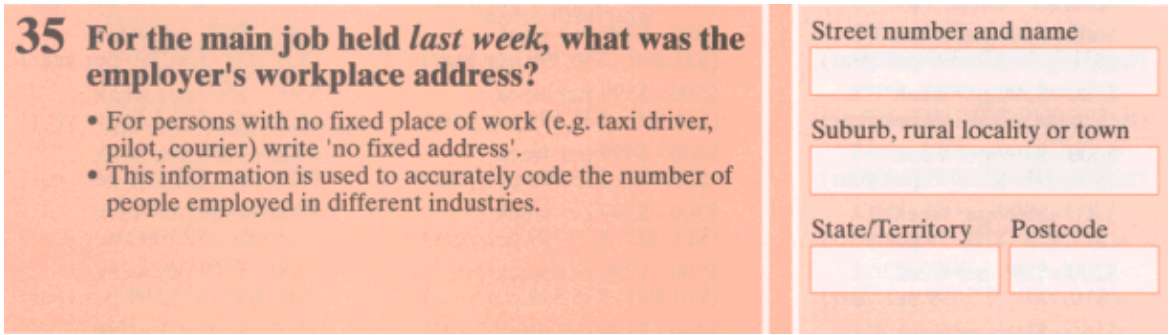
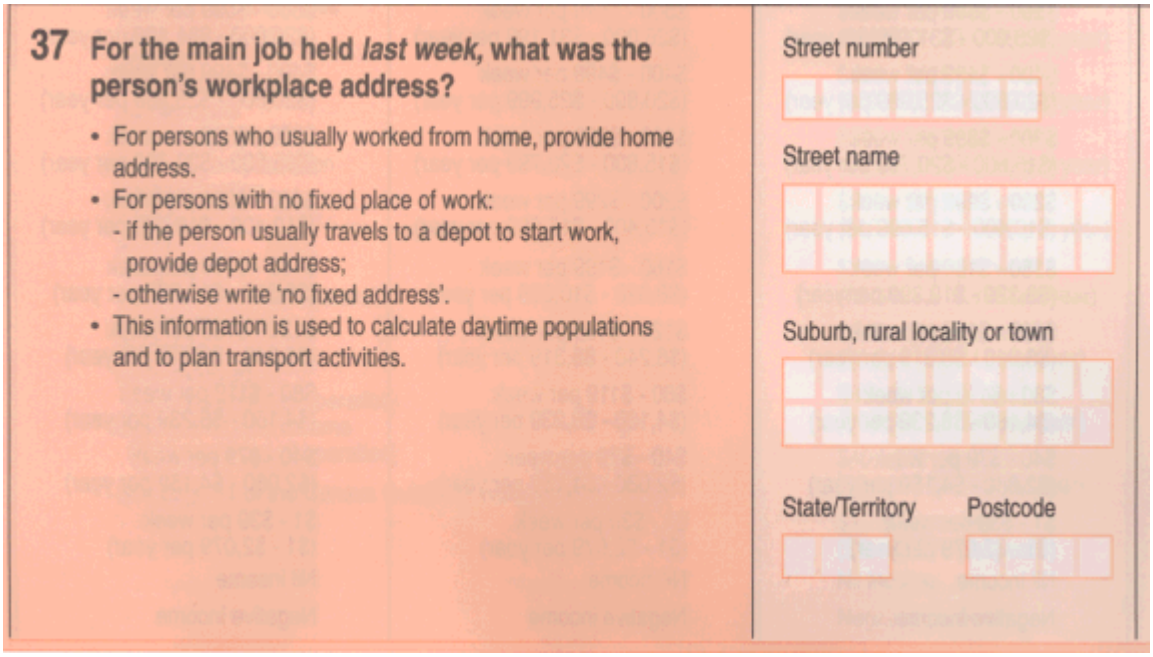


FIGURE 4: PERSON'S WORKPLACE ADDRESS QUESTION, 2001 CENSUS HOUSEHOLD FORM



In 2001, two questions, the first sought a description of the business of the employer and the second, asked for the main goods produced or main services provided by the employer's business, replaced a single question in 1996 which sought details of the industry, business or service carried out by the employer. Refer to Figures 5 and 6.

FIGURE 5: BUSINESS DESCRIPTION QUESTION, 1996 CENSUS HOUSEHOLD FORM

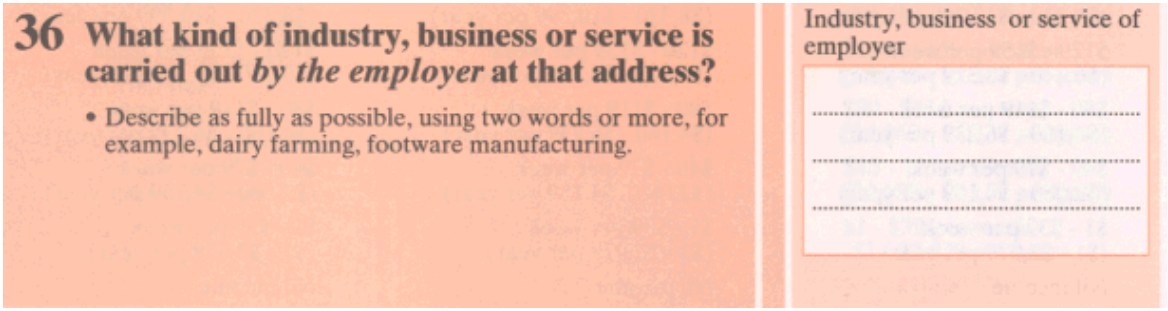
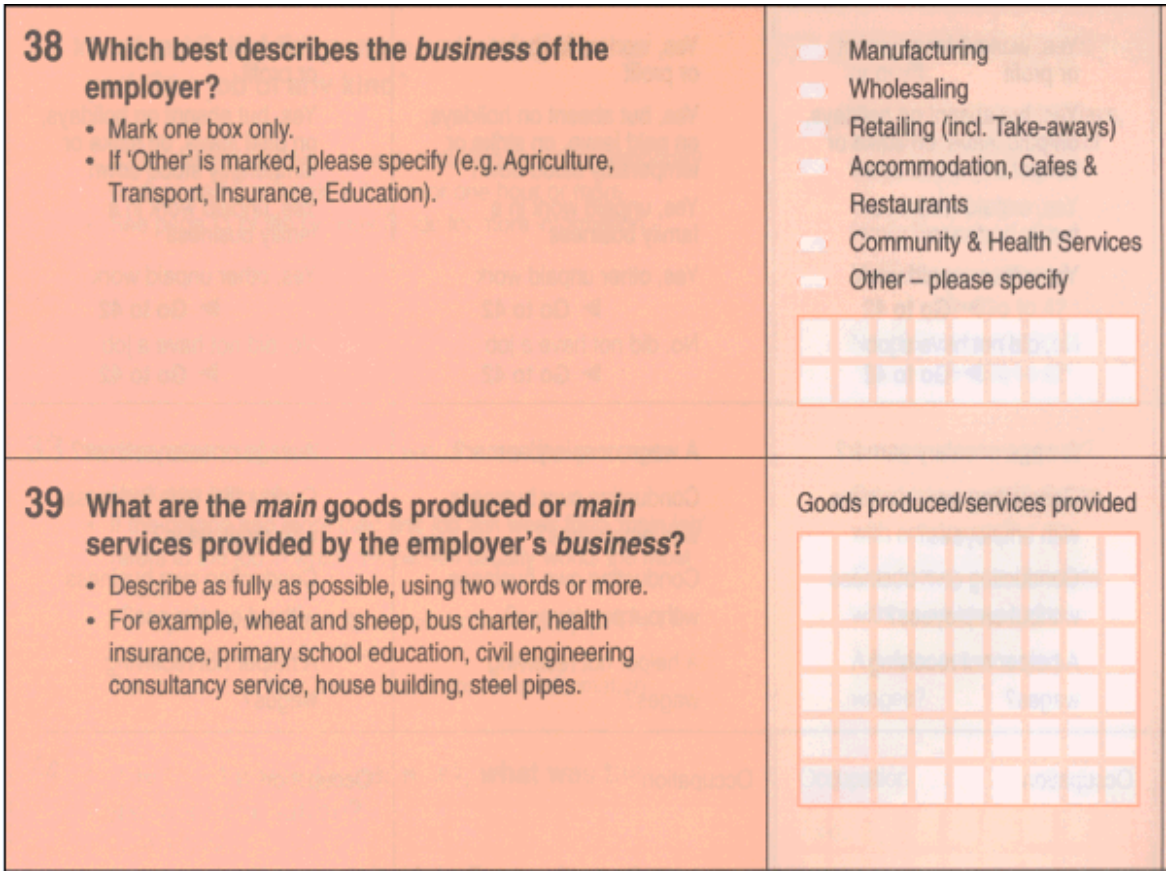


FIGURE 6: BUSINESS DESCRIPTION AND GOODS PRODUCED/SERVICES PROVIDED QUESTIONS, 2001 CENSUS HOUSEHOLD FORM



2.2 The 2001 Census Industry Questions

For the 2001 Census, Industry coding was primarily based on the responses to two questions (Questions 38 and 39). The first question asked for a description of the business of the employer and consisted of a selection of mark-boxes and a write-in section for other responses. The mark-box options had been identified as containing issues that impacted on data quality and Census tests had shown that the use of mark-boxes allowed more accurate classification of the responses.

The second question asked for the main goods produced or main services provided by the employer's business. This question was intended to provide additional information on the activity and the products of the employer's business.

The two questions which preceded the Industry questions, asked the respondent to provide in relation to their main job held last week, their employer's business name and the person's workplace address and are used in some cases to code an outcome. See *Section 4 Processing at the Data Processing Centre (DPC)* for more details about the coding.

The placement of the two labour force questions, including the Full-time/Part-time Job question, the two employment/occupation questions, the questions relating to the employer's name and address and the Industry questions, remained the same as for 1996. The wording of these questions and for the most part the instructions, also remained the same. For further

information about the placement of the labour force questions in relation to Industry response rates, refer to Section 3.1 in Census Working Paper 00/3*: 1996 Census Data Quality: Industry.*

2.2.1 The Full-time/Part-time Job Question

The Full-time/Part-time Job question (Question 32) was the 'gateway' through which respondents answering the Industry questions needed to pass. Four groups of respondents, who answered to Full-time/Part-time Job with:

- Yes, worked for payment or profit;
- Yes, but absent on holidays, on paid leave, on strike or temporarily stood down;
- Yes, unpaid work in a family business; or
- Those who did not respond to the Full-time/Part-time Job question at all,

had their answers to the Industry questions coded.

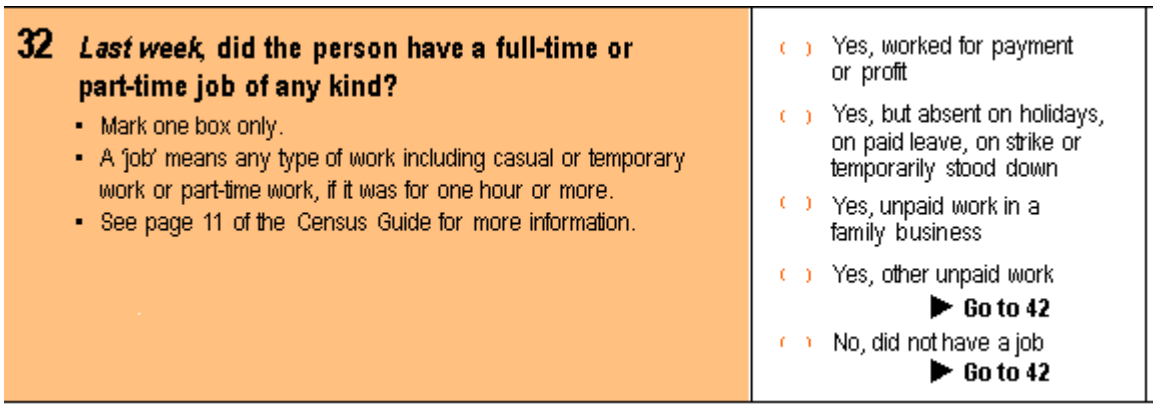
Those who marked the fourth or fifth options:

- Yes, other unpaid work; or
- No, did not have a job,

were sequenced to the Actively Looking for Work question (Question 42), and any responses made to the Industry questions were not coded.

Industry details supplied by respondents who did not answer the 'gateway' question were also coded, to maximise the value of the data.

FIGURE 7: FULL-TIME/PART-TIME JOB (GATEWAY QUESTION), 2001 CENSUS HOUSEHOLD FORM



2.3 Scope of the Industry Questions

The scope of Census Industry data was unchanged from that in 1996. However, in 2001 there were changes to the wording of the questions and the format of response boxes to facilitate ICR technology.

For the 2001 Census and previous censuses, only persons aged 15 years and over, who had a full-time or part-time job of any kind were asked to fill in the Industry questions. Persons were defined as employed if, during the week prior to Census night, they had:

- worked for payment or profit; or
- been absent on holidays, on paid leave, on strike or temporarily stood down; or
- worked as an unpaid worker in a family business.

Industry of employment data were not collected for persons who were unemployed or not in the labour force.

Information in this Paper refers to mainstream enumeration Household and Personal forms and Special Indigenous Personal forms. The following forms are excluded from the analyses of Industry data as respondents did not have the opportunity to answer the Industry questions:

- Substitute forms - used by Census collectors to indicate non-contact with a householder, refusal to submit a Census form, intention to mail-back a form or that a dwelling was unoccupied on Census night.
- Summary forms - used by Census collectors for the enumeration of Non-Private Dwellings where each respondent was given a Personal form.
- Special Short forms - used as part of the Homeless Enumeration Strategy. These forms asked a reduced number of questions to assist in the counting of the homeless who live on the streets as distinct from those living in refuges or permanently living in boarding houses.
- Special Indigenous Household forms - used for the collection of details of the people living and staying in the household and other dwelling related information from people in Indigenous communities. Information was collected mostly by interview.

The Special Indigenous Personal form contained Industry-related questions and as with the mainstream forms, the questions were only asked of employed persons aged 15 years and over. However, interviewers recorded responses to questions on these forms which asked the name and type of the person's job, who they worked for, their workplace address, and what their employer does.

2.4 Relationship Between the Industry and Occupation Variables

There is not necessarily any relationship between an individual's occupation and the Industry in which he or she works. For example, a van driver for an establishment designated as being in the insurance Industry is employed in the insurance Industry and not the transport Industry. Similarly, a teacher at a primary school and a cleaner at a primary school would both be allocated the Industry code 8421 Primary Education. One establishment may employ many people in different occupations but they are all coded to the Industry of the establishment.

The Census recognises this absence of relationship between the Industry and Occupation variables and codes responses to the respective questions separately.

2.5 The Possible Impact of the 'List Effect' on Data

Where a question offers a list of mark-box options for responses, there may be a bias in self-coded responses, known as the 'list effect'.

The impact of this style of question design may include one or more of the following factors:

- an increase in responses to the top option on the list;
- respondents choose a category from the list of response options in preference to one not on the list;
- the response options listed encourage responses different from those that may have been provided without them; and/or
- the options listed influence respondents to answer in a different way, generally in a following write-in section, if available.

During the form design and testing phase of the Census program, questions were assessed for any impact possibly due to 'list effect' before being approved for use in their final format. For more information about the final format of the 2001 Census questions, refer to the Information Paper *2001 Census of Population and Housing: Nature and Content* (cat. no. 2008.0).

The additional Industry question in 2001 (Question 38 on the Household form) incorporated mark-box options for a selection of Industries as well as a write-in section for other responses which together with the question asking for the main goods produced or main services provided by the employer's business (Question 39 on the Household form), was intended to provide additional information on the activity and products of the employer's business.

Responses to Question 38 on the Household form may have been subject to 'list effect' bias.

The following table compares data for the industries listed as mark-box options in the 2001 Census with data obtained in 1996 for those same industries.

TABLE 1: COMPARISON OF 2001 CENSUS INDUSTRY MARK-BOX OPTION RESPONSES WITH 1996 CENSUS RESPONSES

	1996		*2001*		*Intercensal changes*	
	Responses coded to ANZSIC Class (4-digit)	*Responses coded to ANZSIC Division (1- digit)*	*Responses coded to ANZSIC Class (4-digit)*	*Responses coded to ANZSIC Division (1- digit)*	*Responses coded to ANZSIC Class (4-digit)*	*Responses coded to ANZSIC Division (1-digit)*
Mark-box options in 2001 Census (a)	*per cent*		*per cent*		*percentage points*	
Manufacturing	85.7	3.5	83.0	7.0	-2.7	3.5
Wholesaling	88.9	4.6	93.7	4.6	4.8	0.0
Retailing (incl. Takeaways)	95.5	2.6	95.5	2.9	0.0	0.3
Accommodation, Cafes and Restaurants	92.5	. .	98.3	. .	5.8	. .
Community and Health Services	94.3	0.9	86.2	4.8	-8.1	3.9

. . Not applicable. (a) Industry descriptions in the mark-box options list on the 2001 Census form are not necessarily the same as for those in the ANZSIC classification.

In 2001, there were decreases in the percentages of responses coded to the most detailed ANZSIC level in Manufacturing, at the top of the list of options, of 2.7 percentage points, and Health and Community Services, at the bottom of the list, of 8.1 percentage points. Between 1996 and 2001, Manufacturing and Health and Community Services also showed increases of 3.5 percentage points and 3.9 percentage points respectively, in the proportions of employed persons at the ANZSIC division level.

It is not conclusive from the table above how much, if any, intercensal change was attributable to the use of the listed options in the 2001 Census form. The fact that other information supplied on the form was referred to, to assist the classification of Industry, suggests that the 'list effect' may not have had any major impact on Industry coding. See Section 5.4 which discusses the level of response to Questions 38 and 39. See also, *Section 6.2 Undefined Coding Analysis for Industry, 2001 Census*.

3. COLLECTION OF THE DATA

3.1 Enumeration Errors

During the collection phase of the 2001 Census, collectors reported increased difficulty contacting some householders. Access to secure small and large apartment buildings, and gated communities, and growing concerns with regard to security, made it increasingly difficult for collectors to judge whether residents of a building were absent or not. System Created Records (SCRs) were created during Census processing for people for whom a Census form has not been received but where a collector believed that the dwelling was occupied on Census night.

SCRs have values imputed for age, sex, marital status and usual residence only. Values for other variables are set to Not Stated or Not Applicable, depending on the imputed value for age.

An increase in non-response (Not Stated) rates was apparent for many Census variables in the 2001 Census. Most of the change can be attributed to the increase in the proportion of SCRs. A Fact Sheet - *Effect of Census Processes on Non-Response Rates and Person Counts*, has been produced that discusses the factors that may have contributed to the increase in SCRs for 2001 and the percentage of records affected, by state and territory. Please refer to this Fact Sheet on the ABS Website (www.abs.gov.au).

4. PROCESSING AT THE DATA PROCESSING CENTRE (DPC)

4.1 Background to Industry Coding

In 1996, Industry coding to the Australian and New Zealand Standard Industrial Classification (ANZSIC) which provides the framework for classifying statistical units to Industry classes within the ABS, was primarily based on matching employer details with the ABS Business Register (a comprehensive list of Australian businesses coded by Industry classification). Where the coder could not make an appropriate match, a secondary coding process of matching the 'kind of industry, business or service' to a 'simple string' index was used.

Between 1996 and 2001, use of the Business Register was abandoned for Census Industry coding purposes because of the reduced data available, especially at the location level, on the Register. For the 2001 Census, Industry responses were coded by the ABS Coder using the newly developed 'structured' Industry coding index (refer to *Section 4.4.1 The ABS Coder*), making it more consistent with the approaches used to code occupation and qualifications data.

The new two-part question module was also introduced to support the new coding approach.

4.2 Data Capture (DC)

Data Capture (DC) is the process of scanning Census forms into image and text files that are used for all subsequent processes. For the 2001 Census, the Intelligent Character Recognition System (ICR), read hand-written text, verified and corrected the text read from the form, and stored the form image and data for additional processing. Coding staff undertook the 'repair' of information that could not be corrected automatically.

4.3 Stages of Industry Coding

There were three stages to Industry coding. First, Industry codes were automatically allocated by the computer system. Second, codes that could not be allocated automatically, were allocated manually. Third, responses that could not be allocated a code either by Automatic Coding (AC) or Computer Assisted Coding (CAC), were passed on to a Query Resolution (QR) team.

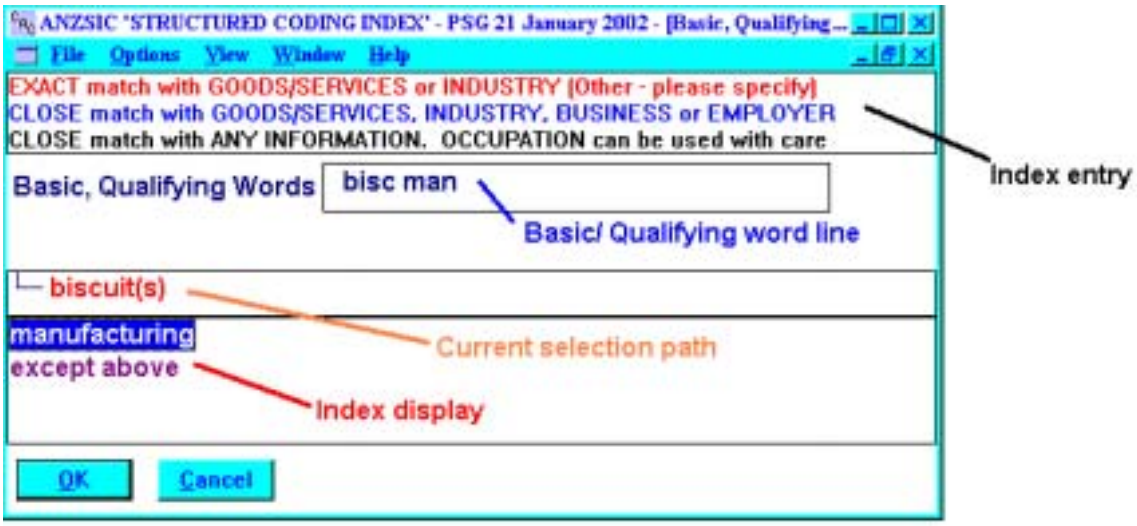
4.3.1 Automatic Coding (AC)

The first stage involved the AC system allocating an Industry code by matching on an index entry, using in the first instance, information from the Goods and Services response (Question 39) for detail, and if necessary, information from the mark-box and business name fields. Where the AC system was unable to allocate an Industry code, coding staff were required to undertake CAC coding.

4.3.2 Computer Assisted Coding (CAC)

The CAC system provided fast access to the Industry coding index. Figure 8 below shows an example of a CAC coding screen for Industry coding.

FIGURE 8: AN EXAMPLE OF A CAC SCREEN FOR INDUSTRY CODING, 2001 CENSUS



The CAC system comprised:

> **Title Line (Basic, Qualifying Word)** - the basic word for ANZSIC coding is a single word which can stand alone as the object of the response provided. A qualifying word in a response identifies an action performed on the object (the basic word).
> **Index Display** - for selecting one or more index entries.
> **Selection Path** - displays which items have been selected in coding the information.
> **Index Entry** - the top index entry (exact match) can only be selected if the basic word of the response and the index display are the same.

If necessary, further lists were presented until a code was determined, or if a system message displayed 'Coding Attempt Unsuccessful', the coder was then instructed to raise a query. Unresolved queries were passed over to the QR team.

Industry information was primarily obtained from the Goods and Services field, that is a response to Question 39 and the Industry mark-box, that is a response to Question 38. Business name and occupation information was used in some cases to assist in getting a correct match.

The CAC operator followed a predefined set of procedures. The two basic elements needed in order to code responses to ANZSIC were the Industry information from the Census forms and an Industry coding system to interpret the information.

4.3.3 Query Resolution (QR)

When the system message 'Coding Attempt Unsuccessful' was displayed, the response was referred to the QR team, who had a wider range of coding tools to assist them in resolving a match. Initially, CAC coding was duplicated and, if unsuccessful, other approaches were attempted, including the use of synonyms, the ANZSIC 'string'-based industry coding index and the ANZSIC structure. Although a match was not always successful, it was a indicator of the quality of the CAC output.

However, a higher than expected initial failure rate for AC and CAC, primarily due to large numbers of vague and incomplete responses, resulted in a higher than expected query rate. Changes were made to the CAC coding methodology following testing of the modified Inteframe Coder (Refer to *Section 4.4.2 The Modified Inteframe Coder*) had revealed that up to a third of incoming queries could be successfully coded using the modified Inteframe Coder as a first step.

If the QR coder could still not allocate an Industry code because the response contained insufficient information, the outcome was 'classified' as a Non-classifiable Economic Unit.

4.4 The Industry Classification and Indexes

The aim of Industry coding is to assign a code for each employed person who has indicated the Industry of the employer for whom they work. The Industry classification used as the basis for coding Industry in the 2001 Census, was the Australian and New Zealand Standard Industrial Classification (ANZSIC).

ANZSIC has a structure comprising four levels: Divisions (the broadest level), Subdivisions, Groups, and Classes (the finest level). At the broadest level, the main purpose is to provide a limited number of categories that will provide a broad overall picture of the economy. For an example of the ANZSIC structure refer to Appendix 1 and for the full classification, the *2001 Census Dictionary*, (cat. no. 2901.0).

For the 2001 Census, Industry responses were coded by the ABS Coder using the newly developed 'structured' Industry coding index. When this was unsuccessful the modified Inteframe Coder was used in an attempt to allocate a code.

4.4.1 The ABS Coder

The ABS uses a coding package/program, commonly known as the ABS Coder, to process responses gathered in censuses and surveys. The coding package/program 'calls on' a specific index, depending on the subject matter being coded, to allocate an appropriate code. In Occupation and Qualification coding, the coder accesses a 'structured' coding index, whereas in past censuses and until recently in ABS surveys, Industry coding was done via a 'string'-based index. For example,

> 'structured' index entry - 2534 acid(s), manufacturing/ acetylsalicylic
>
> 'string'-based index entry - 2534 acetylsalicylic acid mfg.

The information in both entries is the same. However, it is entered differently into the coder. Both indexes continue to be developed by the ABS. See *Section 1.2 Background*.

In preparation for the 2001 Census the 'structured' Industry coding index was tested to determine if it offered any benefits over the 'string'-based Industry coding index. As part of the Census test conducted in September 1998, Census processing staff used the 'structured' Industry coding index which appeared to be of some advantage, as these staff also utilised 'structured' coding indexes when classifying Occupation and Qualification responses. Structured Occupation and Qualification coding indexes were used in the 1996 Census.

It was thought that if all coding indexes were of a structured type then coding staff training, and more generally, their learning requirements, would be reduced, as the skills learnt using Occupation and Qualification coding indexes, could be applied to Industry coding. Therefore, coding staff could easily make the transition from each of the topics using a 'structured' coding index.

Whereas the 'string'-based Industry coding index encouraged, and in fact required, users to think more about their decision than if they were using the 'structured' Industry coding index, the 'structured' coding index aimed to lead users to the correct class by presenting activity listings from which they could make a choice. This step-by-step approach was intended to find the answer without needing to know the title of the Industry class or where it lies within the classification hierarchy. This was expected to promote quick coding, but results of pre-2001 Census tests indicated that any advantage in this area was only marginal. A second, more pronounced, advantage was the reduction of coding inconsistencies which may have been introduced by individual coders through their varying levels of knowledge and different attitudes.

4.4.2 The Modified Inteframe Coder

No Business Register matching for Industry coding was performed in the pre-2001 Census tests despite the fact that in previous censuses Business Register matching accounted for around 50 per cent of the codes allocated, with the remaining 50 per cent by CAC. It was expected that AC would take the place of the Business Register and that CAC would account for the same percentage as coded in the past. However, as coding of test data proceeded, it became evident that a business name would be useful to be certain of a correct code.

For processing of 2001 data, names of some well-known businesses, which employ large numbers of the workforce such as Coles, Telstra and banks and government departments, were added to the 'structured' Industry coding index used by the ABS Coder, but as processing proceeded and query rates rose, a further tool, a modified Inteframe Coder, an index of business names with, in many cases, locality or address information, was provided to the coding teams. In a limited way, business name matching was reintroduced to assist with non-matches.

The modified Inteframe Coder did not contain a full listing of businesses residing on the ABS Inteframe database. Many large businesses were omitted because the decision to use this tool was taken during processing and the DPC did not have the time or resources available to integrate some of the more complex business structures on the Inteframe database into the DPC processing environment.

4.5 *Summary of Industry Coding Methodologies, 1996 and 2001 Censuses*

The different processes used for the 1996 and 2001 Censuses are summarised as follows:

TABLE 2: SUMMARY OF CODING METHODOLOGIES, 1996 AND 2001 CENSUSES

Year	Process	Tools Used	Details
1996	Primary	Business Register	Used employer name and address linked to appropriate ANZSIC code.
	Secondary	ABS Coder - using 'string'-based Industry coding index (a list of goods and services linked to appropriate ANZSIC codes).	Based on response to Question 36 (Industry, Business or Service of Employer).
2001	AC	ABS Coder - using 'structured' Industry coding index (a list of goods and services linked to appropriate ANZSIC codes).	1. Always used Question 39 (Goods and Services) information first. 2. Then, Question 38 (Description of Business) information (mark-box or Other). 3. Then, Question 36 (Business Name). 4. Then, Question 34 (Occupation Title). 5. Then, Question 33 (Own Business values only).
	CAC	ABS Coder - using 'structured' Industry coding index.	1. Used Basic words based on Goods and Services reported. 2. Then, Qualifying words based on Goods and Services provided. 3. Then, Business Name.
	Revised CAC	ABS Coder - using 'structured' Industry coding index, and the modified Inteframe Coder.	The introduction of the modified Inteframe Coder provided an index of business names and localities, linked to ANZSIC.
	QR: Primary Secondary	CAC coding procedures Synonyms, ABS Coder - using 'structured' Industry coding index, and ANZSIC Classification.	Attempting to recode using CAC coding procedures was not always successful, but did provide a quality check for CAC output.
	Revised QR	The modified Inteframe Coder, ABS Coder - using 'structured' Industry coding index.	A stopgap measure until the introduction of the modified Inteframe Coder to the CAC stage.

4.6 Comparison of AC and CAC Coding Rates

Because AC was only introduced in the 2001 Census, it is not possible to provide comparative data with any previous censuses.

A comparison of AC coding rates for ANZSIC Divisions for the 2001 Census is shown in Table 3:

TABLE 3: CODING RATES BY INDUSTRY DIVISION, 2001 CENSUS

	Automatically Coded (AC)		*Not AC'd*	
ANZSIC division	*number*	*per cent*	*number*	*per cent*
Agriculture, Forestry and Fishing	179,626	54.3	151,156	45.7
Mining	39,576	52.6	35,602	47.4
Manufacturing	441,240	43.7	568,939	56.3
Electricity, Gas and Water Supply	34,836	57.4	25,856	42.6
Construction	325,301	58.2	233,281	41.8
Wholesale Trade	210,085	48.1	227,049	51.9
Retail Trade	802,468	66.2	408,864	33.8
Accommodation, Cafes and Restaurants	294,377	71.7	116,212	28.3
Transport and Storage	180,149	50.6	175,725	49.4
Communication Services	101,658	68.5	46,822	31.5
Finance and Insurance	211,601	67.7	100,795	32.3
Property and Business Services	453,174	49.2	467,157	50.8
Government Administration and Defence	169,231	45.8	200,624	54.2
Education	459,404	77.2	135,994	22.8
Health and Community Services	377,105	46.8	429,066	53.2
Cultural and Recreational Services	103,381	51.1	99,075	48.9
Personal and Other Services	188,284	62.6	112,374	37.4
Non-classifiable Economic Units	349	0.7	47,557	99.3
Not stated	0	0.0	144,613	100.0
Total	**4,571,845**	**55.1**	**3,726,761**	**44.9**

An average 55 per cent of responses were coded by the AC system, leaving nearly 45 per cent processed by other means including CAC, QR and Main Edits.

Education had the highest AC rate (77.2 per cent), the next highest was Accommodation, Cafes and Restaurants (71.7 per cent), while Manufacturing had the lowest (43.7 per cent) and Government Administration and Defence (45.8 per cent), the next lowest. Five out of 17 Industry Divisions had AC match rates of less than 50 per cent.

For an examination of the impact of the use of the 'structured' Industry coding index and modified Inteframe Coder on the assignment of Industry codes refer to the Data Quality Investigation (DQI) which used a sample of Collection Districts (CDs) outlined in *Section 5 Sample Data Analysis*.

4.6.1 The Modified Inteframe Coder versus AC and CAC

Codes residing on the modified Inteframe Coder have been determined by contact with the owner or accountant of the business and are based on financial records, whereas information

processed using the ‘structured’ Industry coding index is based on the respondent's description of what main activity takes place at their employer's business. Inconsistencies are inevitably going to occur between the code arrived at using the modified Inteframe Coder and a code arrived at using ‘structured’ Industry coding index, unless there is a business name attached to every entry in the Industry coding index.

4.7 Edits Applied to the Data

The ABS Census program has a minimalist editing approach, with most data output as reported on Census forms. However, editing is the systematic way of altering data to ensure that it is:

- More complete. For example, if the basic demographic variables of age, sex or usual residence are not stated, they are imputed based on known distributions.
- Socially consistent to some extent. For example, age edits do not allow five year olds to be attending high school.
- Consistent with ABS classifications used in other ABS collections. Census Labour Force Status is derived using the same broad derivation used in the Labour Force Survey, to allow clients to more accurately compare data.

There are two key edits applied to Industry data:

1. only persons aged 15 years or over have their Industry details coded, and only if,
2. they answer ‘Yes’ to one of the first three options in the labour force ‘gateway’ question (Question 32 on the Household form) “Last week, did the person have a full-time or part-time job of any kind?”, or did not state an answer to this question.

These edits are entirely logical and should be retained as they comply with standard ABS definitions.

4.8 Explanation of Undefined Coding

The principles of coding to the Australian and New Zealand Standard Industrial Classification (ANZSIC) required responses to be coded to the most detailed level of the classification possible. If a response was not detailed enough to allow coding to the 4-digit level, an undefined code was allocated. The coding structure was:

- The Industry class, or 4-digit level (for example, 7411 for Life Insurance).
- The Industry group, or 3-digit level (for example, 741 for Life Insurance and Superannuation Funds, undefined).
- The Industry subdivision, or 2-digit level (for example, 74 for Insurance, undefined).
- The Industry division, or 1-digit level (for example, K for Finance and Insurance, undefined).

There were three major reasons why undefined coding occurred:

1. Lack of sufficiently detailed information from respondents.
2. The nature and structure of ANZSIC. Some divisions are highly detailed and require precise information from respondents to distinguish one Industry class from another, while other divisions have few entries and coding at the class level can be undertaken with the most basic information.
3. Failure to follow coding procedures rigorously.

Refer to Sections 6.2 and 6.3 for analyses of undefined Industry coding.

4.9 Quality Management and Discrepancy Rates

4.9.1 The Quality Management System

A Quality Management (QM) system was established to identify coding discrepancies, provide feedback to coders and analyse discrepancy rates by topic.

During processing the QM system allowed for the detection of discrepancies and the calculation of a crude discrepancy rate. This crude discrepancy rate differs from a true discrepancy rate for the following reasons:

- A higher proportion of 'poor' coders' work was included in the quality monitoring sample.
- The QM check coders could make the same mistake as the original coder, therefore, the error would not be detected.
- There is not always an absolutely correct code for every response.
- Discrepancies were recorded for any difference between the QM coder although discrepancies at Industry division level were clearly more serious than those at class level. For example, coding Primary Education (8421) to Secondary Education (8422) was given the same weight as coding the Industry division Manufacturing to the Industry division Mining.

The quality of coding using the 'structured' Industry coding index was affected by the following:

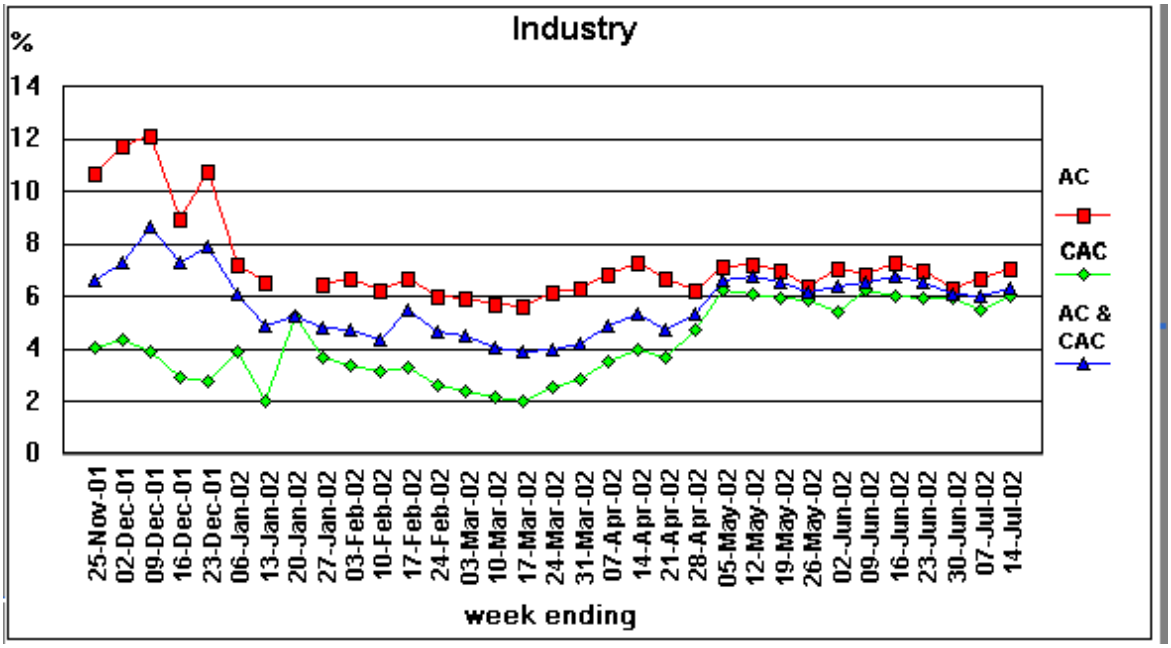
- Information provided by the respondent on the form.
- Training of coding staff.
- Tools available to coding staff.
- Processing methodology changes.

During the processing of the 2001 Census data, a sample of each coder's work was selected for reprocessing by another coder and mismatches were then looked at by an Adjudicator who would decide on the correct code. If the Adjudicator disagreed with the initial coder, a discrepancy would be recorded. There were 8,298,606 applicable Industry counts from which 1,355,093 responses (16.3 per cent) were recorded by QM coders. Altogether, 70,465 Industry discrepancies (5.2 per cent) were recorded in the Management Information System (MIS) reports.

4.9.2 Discrepancy Rates for Industry

Figure 9 below shows the discrepancy rates for Industry over the processing period.

FIGURE 9: DISCREPANCY RATES FOR INDUSTRY, 2001 CENSUS



F

luctuations in the first few months of coding were due to the limited size of the sample as there was only a small amount of Second Release Processing (SRP) coding, which included Industry coding, underway until the end of January 2002.

The initial weeks saw high rates, particularly for AC, as the system was 'bedded down' and systemic AC problems were resolved either through blocking of the AC option or repair of particular letter combinations.

As some previously AC'ed combinations were forced to CAC, the latter's rate rose once more, only to be reduced with time and experience, until coders were encouraged to reduce their frequency of raising queries and to attempt to code to the most detailed level possible.

A new coding facility, a modified Inteframe Coder was introduced into CAC coding in late April 2002 resulting in a slight increase in the discrepancy rate over the following weeks as experience with the new procedures was gained. As Industry coding progressed, the modified Inteframe Coder enabled coders to get a better code and helped reduce the number of Industry responses going to query. See *Section 4.4.2 The Modified Inteframe Coder*.

4.9.3 Discrepancy Rates by Processing Type

There was an expectation that there would be a number of discrepancies between AC and CAC treatment of Subdivisions 41 and 42 because the rules across the two processes were inconsistent, particularly at the start of coding. AC used occupation information if the respondent was self-employed, where CAC did not do this at all. Subsequent changes to the

CAC index and coding screen (which allowed coders to receive a message that the respondent was self-employed and therefore the use of occupation information was appropriate) achieved a marked decline in these discrepancies. However, the changes did not work as fully as intended. Further improvements to the index set-up and consistent paths for the two processes are required before consistent codes across AC and CAC can be achieved.

The nature of the significant AC discrepancies were:

- Codes within Subdivision 42 Construction Trade Services allocated by AC which Adjudicators determined should have been within Subdivision 41 General Construction. For example, a correct code of 4111 House Construction was determined where AC had coded these cases to 4242 Carpentry Services, 4222 Bricklaying Services and 4241 Plastering and Ceiling Services.
- Codes within Subdivision 57 Accommodation, Cafes and Restaurants allocated by AC which Adjudicators determined were incorrect. For example, a correct code of 5720 Pubs, Taverns and Bars was determined where AC had coded these cases to 5710 Accommodation and 5730 Cafes and Restaurants.
- Other discrepancies were as a result of Adjudicators determining that a query should have been raised while AC obtained a code.

The nature of the significant CAC discrepancies were:

- Mostly due to Adjudicators determining that a query should have been raised while coders actually obtained a code.
- Codes within the Subdivision 81 Government Administration and other 8 codes in Divisions N Education and O Health and Community Services, allocated by coders which Adjudicators determined should have been coded to other codes in Subdivision 81 Government Administration. For example, a correct code of 8112 State Government Administration was determined where coders had coded these cases to 8111 Central Government Administration, Subdivision 84 Education and 8420 School Education.
- Codes within the Subdivision 84 Education and other 8 codes in Divisions N Education and O Health and Community Services, allocated by coders which Adjudicators determined should have been coded to other codes in Subdivision 84 Education. For example, a correct code of 8420 School Education was determined where coders had coded these cases to Subdivision 84 Education, 8111 Central Government Administration and 8112 State Government Administration.

4.10 Validation

The role of validation in the processing system was to ensure that the data produced, and released, met the requirements of users. This role was carried out by checking the data produced by the system to ensure that it met the stated output requirements, and identifying and correcting, the errors that occurred. When the source of an error was identified, the part of the system that was generating the error was reviewed for the most suitable method of correction. In some cases, a procedural correction was more appropriate than a system update.

5. SAMPLE DATA ANALYSIS

5.1 Data Quality Investigation (DQI) Sample

A 2 per cent statistically derived sample of Collection Districts (CDs), numbering approximately 740, was taken for detailed quality analysis. Included in the sample were CDs from each state and territory representing the wide range of urban and rural areas in Australia.
Using this sample, Data Quality Investigation (DQI) tasks, directly related to the areas for which in-depth investigations were planned, were carried out by a DQI team at the Data Processing Centre (DPC). The resulting data quality information is made available to clients in Census Papers and other related publications.

5.2 Comparison of the Modified Inteframe Coder and the 'Structured' Industry Coding Index

The processing of Industry data has changed considerably since the 1996 Census in the following ways:

- In 1996 the Business Register was used to code Industry according to, in the first instance, employer name and address details, followed by an attempt to code using a 'string'-based coding index, if the Business Register could not find a match.
- In 2001 AC and CAC used a 'structured' Industry coding index. Initially this was done using the modified Inteframe Coder as a secondary measure.
- In 2001 coding was based on responses to two questions rather than one, as was the case in 1996. The first question but second coding step, being a mark-box or write-in description of the business of the employer and the second question but first coding step, a write-in description of the main goods produced or services provided by the employer's business.

In addition to these major changes, several minor procedural changes occurred during the processing cycle. The combination of the above changes and the structural changes taking place in the economy make it difficult to quantify and apportion the degree to which each processing change was attributable.

The DQI team investigated the impact on Industry data of the use of the 'structured' Industry coding index compared with the assignment of Industry codes using the workplace address question by coding via the modified Inteframe Coder. However, only broad indications about the effects on the changes were produced for the reasons mentioned above.

5.2.1 Using the Modified Inteframe Coder to Obtain an ANZSIC Code

Utilising a maximum timeframe of 60 seconds per record to try the standard and lateral searches to achieve an Industry class match, the DQI team obtained Industry codes using the modified Inteframe Coder as shown in the following table:

TABLE 4: INDUSTRY CODES OBTAINED USING THE MODIFIED INTEFRAME CODER, 2001 DQI SAMPLE

Component details	*Number of Persons*	*Per cent*
Total (a)	**369,456**	
Not Applicable (NA)	191,833	51.9
Not Stated (NS)	15,338	4.2
Total excl. NA and NS	162,285	43.9
ANZSIC obtained	49,850	30.7
ANZSIC not obtained	112,435	69.3

(a) Includes overseas visitors.

The low match rate of 30.7 per cent in the sample is attributed to problems either relating to responses in the Census forms such as the provision of an incomplete business name, incorrect spelling of the business name, provision of a brand name rather than a trading name, or shortcomings associated with the modified Inteframe Coder, such as the business location in the Coder not matching that given by the respondent, incomplete listing of businesses (both large and small) and errors in business names.

The addition of postcodes as a field could give the coder greater discretion in cases where an exact match is not possible, but other available information suggests that a match is likely. A complete list of Inteframe units, groomed to allow for automatic repair issues likely to arise, and the requirement to code to location level, would further assist in achieving a higher match rate.

5.2.2 Using the 'Structured' Industry Coding Index to Obtain an ANZSIC Code

Coders were trained in the use of the ABS CAC Coder using the 'structured' Industry coding index to obtain Industry codes. The sample was limited to that for which the DQI team had obtained an ANZSIC code using the modified Inteframe Coder and was taken when the 2 per cent sample was approximately 96 per cent complete, due to time constraints. Therefore, instead of a starting total of 49,850 as shown in Table 4, the total used was 42,755. The results are shown in the following table:

TABLE 5: INDUSTRY CODES OBTAINED USING 'STRUCTURED' INDUSTRY CODING INDEX, 2001 DQI SAMPLE

Component details	*Number of Persons*	*Per cent*
Total	**42,755**	
Not Stated (NS)	92	0.2
Total excl. NS	42,663	99.8
ANZSIC obtained	34,133	80.0
ANZSIC not obtained	8,530	20.0

The 80 per cent match rate may be artificially high, due to the nature of the sample used. Given that the 42,755 records in the 'structured' Industry coding index sample had previously been successfully coded using the modified Inteframe Coder, they were likely to be higher quality records, thus facilitating the next phase, coding by the 'structured' Industry coding index.

5.2.3 Results of Comparison of the Modified Inteframe Coder and the 'Structured' Industry Coding Index

It must be recognised that output obtained coding via the modified Inteframe Coder and a 'structured' Industry coding index is by two very different coding methodologies. Codes residing on the modified Inteframe Coder are determined by contact with the owner or accountant of the business and are based on financial records, whereas information used with the 'structured' Industry coding index is provided by an employee based on a description of what 'happens' at their place of work. Therefore, inconsistencies are likely to occur when comparing a code arrived at via the modified Inteframe Coder and using the 'structured' Industry coding index, unless for every entry in the 'structured' Industry coding index, a business name is attached.

For the 34,133 records for which an ANZSIC code was obtained using both the modified Inteframe Coder and the 'structured' Industry coding index, the resultant codes were as follows:

TABLE 6: COMPARISON OF RESULTS OBTAINED FROM THE MODIFIED INTEFRAME CODER AND THE 'STRUCTURED' INDUSTRY CODING INDEX, 2001 DQI SAMPLE

Component details	*Frequency*	*Per cent*
No match at any level	8,025	23.5
Division level match	26,108	76.5
Subdivision level match	23,871	69.9
Group level match	21,258	62.3
Class level match	14,738	43.2
Total	**34,133**	**100.0**

Using only the modified Inteframe Coder to obtain the ANZSIC codes the success rate was 30.7 per cent, whereas using only the 'structured' Industry coding index, the rate was 80.0 per cent. A comparison of the two methods showed that 76.5 per cent matched at the Division level, 69.9 per cent matched at the Subdivision level, 62.3 per cent matched at the Group level and 43.2 per matched at the Class level.

Of the 8,025 'no match' records, using the modified Inteframe Coder, ANSZIC Divisions Manufacturing occurred in 15 per cent, Wholesale Trade in 15 per cent and Property and Business Services in 20 per cent of the cases. Using the 'structured' Industry coding index, ANZSIC Divisions Manufacturing occurred in 18 per cent and Retail Trade in 15 per cent of the cases.

The most common modified Inteframe Coder and 'structured' Industry coding index discrepancies were in the ANZSIC Divisions in the following table:

TABLE 7: MOST COMMON MODIFIED INTEFRAME CODER AND 'STRUCTURED' INDUSTRY CODING INDEX DISCREPANCIES, 2001 DQI SAMPLE

ANZSIC Division coded by modified Inteframe Coder	*ANZSIC Division coded by 'structured' Industry coding index*	*Discrepancy Rate Per cent*
Wholesale Trade	Manufacturing	6.3
Wholesale Trade	Retail Trade	4.9
Manufacturing	Transport and Storage	3.9
Property and Business Services	Manufacturing	3.4
Manufacturing	Communication Services	3.2

The Manufacturing Industry division appeared to have the highest rate of discrepancies for both the modified Inteframe Coder and the 'structured' Industry coding index.

5.3 Industry from Business Name versus Industry from Mark-box Question

A second investigation looked at the correlation between what the respondent answered for Question 38 (the mark-box Industry question) and the Industry that the DQI coders were able to code the related responses to business name and workplace address (Questions 36 and 37), to.

The following table shows the correlation between respondents' answers and the DQI coders' 'matches':

TABLE 8: CORRELATION BETWEEN RESPONDENTS' ANSWERS TO THE MARK-BOX INDUSTRY QUESTION AND CODERS' 'MATCHES' USING RESPONSES TO THE BUSINESS NAME AND WORKPLACE ADDRESS QUESTIONS, 2001 DQI SAMPLE

	Business name and address Questions 36 and 37						
Mark-box Question 38	*Manufact-uring.*	*Wholesale Trade*	*Retail Trade*	*Accomm., Cafes and Restaur-ants*	*Health and Communit y Services*	*Other stated, and matched*	*Total matched responses*
Manufacturing	4,818	665	244	12	39	1,196	6,974
Wholesaling	246	992	240	10	5	346	1,839
Retailing (incl. Take-aways)	178	675	4,896	280	52	1,190	7,271
Accomm., Cafes & Restaurants	16	11	353	1,707	39	233	2,359
Community & Health Services	49	20	144	61	2,865	1,702	4,841
Other marked	8	5	10	4	26	128	181
Write-in	517	297	469	330	282	5,720	7,615
Combination responses (a)	1,214	696	923	557	538	13,280	17,208
Not stated	154	66	165	65	60	566	1,076
Total	**7,200**	**3,427**	**7,444**	**3,026**	**3,906**	**24,361**	**49,364**

(a) Combination (multi-mark) response examples include: Manufacturing + Wholesaling, Manufacturing + write in, Manufacturing + Other + write-in.

Table 8 shows that 69.1 per cent of respondents who marked the Manufacturing box in the mark-box Industry question (Question 38) were subsequently matched by DQI coders to the Manufacturing division using responses to the business name and workplace address questions (Questions 36 and 37). 53.9 per cent of those who marked the Wholesaling box were matched to the Wholesale Trade division; 67.3 per cent of those who marked the Retailing box were matched to the Retail Trade division; 72.4 per cent of those who marked the Accommodation, Cafes and Restaurants box were matched to the Accommodation, Cafes and Restaurants division, and 59.2 per cent of those who marked the Community and Health Services box were matched to the Health and Community Services division.

In all of the above cases the correlation between Question 38 and Questions 36 and 37 was greater than 50 per cent but given the reported limitations of DPC matching using the modified Inteframe Coder, it is possible that the correlation was even greater.

5.4 Completion of the Two Industry Questions

An examination of the DQI Industry data, cross-tabulated with the Labour Force sequencing question (Question 32) - ' *Last week*, did the person have a full-time or part-time job of any kind?'- identified that over 99.4 per cent of persons who indicated that they had a job during the previous week, gave some information about their employer's business activity as shown in Table 9 below.

Of the 165,640 people who responded to the Industry questions, 99.0 per cent answered the mark-box question (Question 38), while 94.8 per cent answered the Goods and Services question (Question 39). 4.3 per cent (7,079) answered the mark-box question only, while even less (1,597) answered the Goods and Services question only.

Overall, 93 per cent of the maximum number of persons eligible to answer the Industry questions provided the information appropriately by giving a valid response to the mark-box question, plus a description of the main goods and services provided by their employer. However, as information provided in either question could ultimately be used to classify the business to a particular Industry, 98.7 per cent (163,507 persons) in the DQI sample provided valid information for Industry coding purposes.

TABLE 9: RESPONSE RATES TO INDUSTRY QUESTIONS, 2001 DQI SAMPLE

	Number	*Per cent*
Relevant populations:		
Total in DQI sample (excluding Overseas visitors)	366,667	
Total persons with a single or multi-marked responses to the Labour Force sequencing question, where one (or more) of the first three options was marked (maximum in-scope population for Industry questions)	166,648	
Response to Question 38:		
Persons who answered Question 38 (single valid mark or 'Other + write-in'):		
and Question 39	154,925	
but not Question 39	6,985	
Persons who multi-marked Question 38:		
and answered Question 39	2,039	
but did not answer Question 39	94	
Total persons who answered Question 38:		
and Question 39	156,964	94.8
but not Question 39	7,079	4.3
Total persons who answered Question 38	164,043	99.0
Total persons who answered Question 39, but not Question 38	1,597	1.0
Total persons who responded to an Industry question	**165,640**	

Two-thirds of the 87,680 people in the sample who reported an Industry at Question 38, other than Manufacturing, Wholesaling, Retailing, Accommodation, Cafes and Restaurants or Community and Health Services, marked the 'Other' box and then filled in the write-in boxes. However, a further 32.9 per cent of respondents ignored marking the 'Other' box at all, proceeding to the write-in boxes below to supply details of their employer's business. The resulting high level of completion suggests that it may not be necessary to have a two-stage process (i.e. an 'Other' mark box, plus write-in boxes) to elicit such information. It is recommended that the requirement for the 'Other- please specify' for this, and similar mark-box questions on the form, is tested before the next Census.

The omission of a mark-box for 'Other - please specify' will also eliminate the occurrence of people marking 'Other' but not supplying further business information (as happened with 1,058 people in the sample).

5.5 Multiple Marks

Response options for the question 'Which best describes the *business* of the employer?' (Question 38) included a selection of mark-boxes, and an 'Other - please specify', plus write-in combination. In the DQI sample, 164,043 people responded to Question 38. Of these, 98.7 per cent marked just one box (or the 'Other' plus write-in section), while 1.3 per cent marked more than one response.

The most common mark-box only combinations were: Wholesaling plus Retailing (194 responses); Manufacturing plus Wholesaling (167 responses), and Manufacturing plus Retailing (107 responses).

However, the majority of the multiple marks for Question 38 included a combination of Industry mark-box plus a written description of the Industry in the write-in box. Of the 2,133 people who multi-marked Question 38, 1,378 (64.6 per cent) used the write-in box to supply extra information. (See Table 10.) This additional information was unexpected, as previous studies have shown higher response rates for (simple) mark-box formats, than for questions requiring (more complex) text answers.

TABLE 10: MULTIPLE MARKING OF INDUSTRY MARK-BOX QUESTION, INCLUDING A WRITE-IN RESPONSE, 2001 DQI SAMPLE

	Total	*Per cent*
Manufacturing + write-in	233	16.9
Manufacturing + (Other + write-in)	81	5.9
Wholesaling + write-in	76	5.5
Wholesaling + (Other + write-in)	29	2.1
Retailing + write-in	314	22.8
Retailing + (Other + write-in)	108	7.8
Accom. Cafes and Restaurants + write-in	86	6.2
Accom. Cafes and Restaurants + (Other + write-in)	58	4.2
Community and Health Services + write-in	250	18.1
Community and Health Services + (Other + write-in)	92	6.7
Balance of combinations	51	3.7
Total Multiple Marks Including 'Write-in' Combinations	**1,378**	**100.0**

6. FINAL DATA ANALYSIS

The 2001 Census was a self-enumerated questionnaire completed by respondents with little or no assistance from Census collectors. Therefore, data quality relied heavily on the ability of respondents to understand each question and to answer in the appropriate manner with the appropriate amount of detail. It was also crucial to have adequate strategies to process insufficient responses.

6.1 Non-response Rates

The overall non-response rate for Industry of employment deceased slightly from 2.0 per cent in the 1996 Census to 1.7 per cent in the 2001 Census. The maintenance of such an acceptable rate of non-response for 2001 may have been due to the changes in form design with the use of two questions and the mark-box options. In most cases, Industry coding was achieved more accurately and definitively, and a response to either question or a partial response to both questions, could constitute a response.

6.1.1 Non-response Rates, 2001 Census

In 2001, Industry had the third lowest non-response rate, after Occupation and Job Last Week, (with non-response rates of 1.2 per cent and 1.4 per cent respectively), for responses by employed persons aged 15 and over. It should be noted that, similarly to Industry, Occupation had two questions in 2001, giving it an advantage of being coded as Not Stated, only if neither of the two questions was responded to.

The non-response rate for the Industry of employment variable for the 2001 Census compares favourably with the rates for other variables applicable to the employed population aged 15 years or more.

6.1.2 Comparison of Non-response Rates, 1996 and 2001 Censuses

The placement of the labour force questions (including Industry) and their subsequent sequencing remained unchanged for 2001. This overcame the loss of Industry data that had occurred prior to 1996 when an instruction on the form resulted in respondents who had indicated that they were not looking for work, skipping the remaining employment questions. For further information about the placement of, and the wording and instructions for the labour force questions in relation to response rates for Industry, refer to Section 3.1 in *Census Working Paper 00/3: 1996 Census Data Quality: Industry,* and Section 6.2 in *Census Paper No. 03/05 2001 Census: Labour Force Status.*

6.1.3 Characteristics of Non-respondents

TABLE 11: INDUSTRY BY STATED/NOT STATED, BY SEX, AGE, INCOME, OCCUPATION AND BIRTHPLACE, 2001 CENSUS

	Industry			
	Stated		*Not stated*	
Variable	*Number*	*Per cent*	*Number*	*Per cent*
Sex:				
Male	4,470,725	98.3	76,058	1.7
Female	3,683,268	98.2	68,555	1.8
Age:				
15 to 19	533,770	97.0	16,456	3.0
20 to 29	1,768,169	98.4	28,265	1.6
30 to 39	2,001,579	98.6	29,008	1.4
40 to 49	2,044,103	98.6	28,364	1.4
50 to 59	1,398,489	98.5	21,892	1.5
60 to 69	345,058	96.8	11,244	3.2
70 to 79	51,829	88.8	6,525	11.2
80 to 89	8,646	78.0	2,436	22.0
90 to 99	2,113	85.9	348	14.1
100 and over	237	76.0	75	24.0
Income:				
Negative	27,374	95.1	1,414	4.9
Nil	36,622	87.9	5,025	12.1
$1-399	2,086,839	97.4	56,581	2.6
$400-999	4,333,231	98.9	46,314	1.1
$1,000 or more	1,501,180	99.4	8,371	0.6
Not stated	168,747	86.2	26,908	13.8
Occupation:				
Managers and Administrators	760,002	99.4	4,821	0.6
Professionals	1,506,753	99.5	7,343	0.5
Associate Professionals	970,767	99.5	4,886	0.5
Tradespersons and Related Workers	1,008,291	99.0	10,612	1.0
Advanced Clerical and Service Workers	307,512	99.2	2,456	0.8
Intermediate Clerical, Sales and Service Workers	1,355,313	99.2	11,388	0.8
Intermediate Production and Transport Workers	663,874	99.0	6,947	1.0
Elementary Clerical, Sales and Service Workers	785,017	99.1	7,361	0.9
Labourers and Related Workers	705,548	98.3	11,909	1.7
Not stated	26,116	26.4	72,713	76.3
Birthplace:				
Australia	6,061,790	98.4	96,711	1.6
Overseas	1,963,973	98.0	40,405	2.0
Inadequately described	7,206	93.2	522	6.8
Not stated	121,024	94.6	6,975	5.4

As indicated in Table 11, there were no significant differences in the response rates in terms of whether the respondent was male or female.

Persons over the age of 60 were more likely not to state their Industry of employment and the level of non-response increased more significantly for persons over 70 years of age which is consistent with findings in other Census Papers dealing with employment-related variables. The high proportion of Not Stateds may be a function of respondents over the age of 60, not considering that the question was relevant to them and thereafter not responding, instead of marking the 'No, did not have a job' option in the 'gateway' (Full-time/Part-time Job) question.

76.3 per cent of persons who did not state their Industry of employment, also did not state their Occupation.

Non-respondents to Industry were also more likely to have Negative, Nil or Not stated Income.

5.4 per cent of persons who did not state their Industry of employment, also did not state their Birthplace.

6.2 Undefined Coding Analysis for Industry, 2001 Census

Table 12 below, shows the frequency of undefined coding for each ANZSIC division in 2001. Undefined coding percentages in Tables 12, 13 and 14 have been adjusted to eliminate the effects of the structure of ANZSIC on undefined coding rates. For example, Industry codes like the ANZSIC subdivision Rail Transport (62) or the group Fruit and Vegetable Processing (213) represent the most detailed code and are therefore treated in this analysis, as an ANZSIC class rather than as a subdivision or group respectively.

TABLE 12: UNDEFINED CODING RATES BY INDUSTRY DIVISION, 2001 CENSUS

ANZSIC division	*% of responses coded to ANZSIC division (1-digit)*	*% of responses coded to ANZSIC subdivision (2-digit)*	*% of responses coded to ANZSIC group (3-digit)*	*% of responses coded to ANZSIC class (4-digit)*	*Total Persons*
Agriculture, Forestry and Fishing	0.7	3.3	3.1	92.9	330,782
Mining	5.3	2.2	0.9	91.5	75,178
Manufacturing	7.0	4.0	5.9	83.0	1,010,179
Electricity, Gas and Water Supply	0.5	9.8	. .	89.7	60,692
Construction	2.6	1.0	8.2	88.2	558,582
Wholesale Trade	4.6	0.5	1.2	93.7	437,134
Retail Trade	2.9	1.1	0.4	95.5	1,211,332
Accommodation, Cafes and Restaurants	. .	1.7	. .	98.3	410,589
Transport and Storage	6.4	10.7	3.4	79.5	355,874
Communication Services	. .	0.9	. .	99.1	148,480
Finance and Insurance	0.6	4.5	0.5	94.4	312,396
Property and Business Services	0.2	2.3	1.5	96.0	920,331
Government Administration and Defence	0.1	0.1	0.3	99.5	369,855
Education	. .	3.0	2.6	94.5	595,398
Health and Community Services	4.8	7.5	1.4	86.2	806,171
Cultural and Recreational Services	1.9	0.7	1.4	96.0	202,456
Personal and Other Services	. .	0.2	0.3	99.5	300,658
Non-classifiable Economic Units	. .	. .	. .	. .	47,906
Not stated	. .	. .	. .	. .	144,613
Total					**8,298,606**

. . Not applicable.

Table 12 above shows that Transport and Storage contained the highest level of undefined coding, with only 79.5 per cent of the responses in this division coded to the ANZSIC class level. 6.4 per cent of the responses were only able to be coded to the ANZSIC division level and 10.7 per cent only able to be coded to the subdivision level, with most of the undefined coding occurring in the Air and Space Transport (64) subdivision.

Manufacturing contained the second highest level of undefined coding, with only 83.0 per cent of the responses in this division coded to the ANZSIC class level. 7.0 per cent of the responses were only able to be coded to the ANZSIC division level and 5.9 per cent only able to be coded to the group level, with most of the undefined coding occurring in the Clothing Manufacturing (224), Log Sawmilling and Timber Dressing (231) and Plastic Product Manufacturing (256) groups.

Other high levels of undefined coding featured in Health and Community Services and Construction, with only 86.2 per cent and 88.2 per cent respectively, of responses in these divisions coded to the most detailed code. Health and Community Services contained 4.8 per cent of responses coded to the division level and 7.5 per cent coded to the subdivision level. Construction contained 2.6 per cent of responses coded to the division level and 8.2 per cent of responses coded to the group level.

The lowest levels of undefined coding occurred in the Government Administration and Defence and Personal and Other Services industries (both with 99.5 per cent) and Communication with 99.1 per cent of responses in these divisions coded to the ANZSIC class level.

6.3 Undefined Coding Comparison for Industry, 1996 and 2001 Censuses

In the following analysis, 2001 Census ANZSIC undefined coding is compared with 1996 Census ANZSIC undefined coding to identify significant increases or decreases. Table 13 shows the percentage of responses coded at ANZSIC Division, Subdivision, Group and Class level for 1996 and Table 14 shows net changes in undefined coding between 1996 and 2001.

TABLE 13: UNDEFINED CODING RATES BY INDUSTRY DIVISION, 1996 CENSUS

ANZSIC division	*% of responses coded to ANZSIC division (1-digit)*	*% of responses coded to ANZSIC subdivision (2-digit)*	*% of responses coded to ANZSIC group (3-digit)*	*% of responses coded to ANZSIC class (4-digit)*	*Total Persons*
Agriculture, Forestry and Fishing	1.1	22.8	9.4	66.8	324,319
Mining	8.0	13.7	7.8	70.4	86,261
Manufacturing	3.5	3.7	7.1	85.7	965,025
Electricity, Gas and Water Supply	0.3	1.1	. .	98.7	58,698
Construction	6.4	2.6	8.7	82.3	484,078
Wholesale Trade	4.6	0.8	5.6	88.9	446,543
Retail Trade	2.6	1.3	0.6	95.5	1,036,639
Accommodation, Cafes and Restaurants	. .	7.5	. .	92.5	355,283
Transport and Storage	5.0	11.6	3.7	79.7	332,074
Communication Services	. .	1.5	0.2	98.4	150,188
Finance and Insurance	0.1	8.7	. .	91.2	296,453
Property and Business Services	. .	0.8	1.0	98.1	750,185
Government Administration and Defence	0.8	1.1	0.7	97.4	373,422
Education	. .	3.8	4.7	91.5	540,059
Health and Community Services	0.9	3.8	1.0	94.3	725,168
Cultural and Recreational Services	1.1	0.8	2.0	96.2	179,050
Personal and Other Services	. .	0.1	0.1	99.8	277,904
Non-classifiable Economic Units	. .	. .	. .	. .	103,142
Not stated	. .	. .	. .	. .	151,368
Total					**7,635,859**

. . Not applicable.

TABLE 14: LEVEL OF MOVEMENT IN UNDEFINED CODING RATES BY INDUSTRY DIVISION, 1996 AND 2001 CENSUSES

ANZSIC division	*Change in % of responses coded to ANZSIC division (1-digit)*	*Change in % of responses coded to ANZSIC subdivision (2-digit)*	*Change in % of responses coded to ANZSIC group (3-digit)*	*Change in % of responses coded to ANZSIC class (4-digit)*	*Change in Total Persons*
Agriculture, Forestry and Fishing	-0.4	-19.5	-6.3	26.1	6,463
Mining	-2.7	-11.5	-6.9	21.1	-11,083
Manufacturing	3.5	0.3	-1.2	-2.7	45,154
Electricity, Gas and Water Supply	0.2	8.7	0	-9.0	1,994
Construction	-3.8	-1.6	-0.5	5.9	74,504
Wholesale Trade	0	-0.3	-4.4	4.8	9,409
Retail Trade	0.3	-0.2	-0.2	0	174,693
Accommodation, Cafes and Restaurants	0	-5.8	0	5.8	55,306
Transport and Storage	1.4	-0.9	-0.3	-0.2	23,800
Communication Services	0	-0.6	-0.2	-0.7	-1,708
Finance and Insurance	0.5	-4.2	0.5	3.2	15,943
Property and Business Services	0.2	1.5	0.5	-2.1	170,146
Government Administration and Defence	-0.7	-1.0	0.4	2.1	-3,567
Education	0	-0.8	-2.1	3.0	55,339
Health and Community Services	3.9	3.7	0.4	-8.1	81,003
Cultural and Recreation Services	0.8	-0.1	-0.6	-0.2	23,406
Personal and Other Services	0	0.1	0.2	-0.3	22,754
Non-classifiable Economic Units	. .	. .	. .	. .	-55,236
Not stated	. .	. .	. .	. .	-6,755
Total					**662,747**

. . Not applicable.

According to Table 14, the most marked increase between the 1996 and 2001 Censuses in responses coded to the most detailed ANZSIC level occurred in the Agriculture, Forestry and Fishing (up 26.1 percentage points). A large proportion of the increase can be attributed to the fall of 19.5 percentage points in the proportion of responses allocated a subdivision code. In 1996, the high proportion of responses allocated the subdivision code was due to the reliance by coders on the often inadequate description by respondents (e.g. 'Farmer') and the smaller proportion of agricultural businesses on the Business Register which reduced the likelihood of business matching. The improved level of defined coding in 2001 can be attributed to the work by classifications staff on the Agriculture and Mining areas of the coding index.

Mining had the second highest increase (up 21.1 percentage points). Improved specification of the mined product due to form and coding process changes, contributed to the fall in the proportion of responses coded to the 1-digit level.

Other increases occurred in Construction (up 5.9 percentage points), Accommodation, Cafes and Restaurants (up 5.8 percentage points) and Wholesale Trade (up 4.8 percentage points). The divisions of Electricity, Gas and Water Supply with a decrease of 9.0 percentage points, Health and Community Services with a decrease of 8.1 percentage points and Manufacturing and Property and Business Services with smaller decreases of 2.7 and 2.1 percentage points respectively, suggest that there are still problems arising from respondents' insufficient responses to the Industry description question and that the changes to the form and coding process in 2001 have not significantly reduced the level of coding to the 1, 2 and 3-digit level.

Overall, very little change has occurred with the quality of Industry data as measured by the rate of undefined coding data, as a result of the process and form design changes for 2001. Whilst the quality of some Industry divisions has improved, other divisions have decreased in quality.

7. RECONCILIATION OF 2001 CENSUS INDUSTRY DATA WITH AUGUST 2001 LABOUR FORCE SURVEY DATA

7.1 Data Reconciliation Methodology

The purpose of this section is to explain the differences in the collection of Industry data between the Labour Force Survey and the Census, to outline the steps taken to reconcile these two data collections and to present the findings from this reconciliation. The methodology used to reconcile Census and Labour Force Survey data is based on an internal paper called *Comparing Labour Force Survey and Population Census Data*, prepared by the ABS' Labour Force Section and Census Development and Field Organisation Section in January 1998.

Although the Census and the Labour Force Survey both collect data on Industry, they are not strictly comparable due to differences in the scope, coverage, timing, measurement of underlying concepts and collection methodology. Factors contributing to differences in estimates include:

- under-enumeration in the Census for which Census Industry data were not adjusted;
- the use in the Labour Force Survey of population benchmarks derived from incomplete information about population change;
- differing treatments for non-response to the Census and the Survey;
- the personal interview approach adopted in the Survey as opposed to self-enumeration in the Census; and
- sampling variability.

Differences in the underlying definition of 'employed' between the two collections should also be borne in mind when comparing figures. Census questions are not as detailed, nor as comprehensive as the Labour Force Survey questions which is largely due to space limitations on the Census form, as well as constraints imposed by self-enumeration. The differences in definition of 'employed' between the two collections relate specifically to absences from work.

To determine the labour force status of persons absent from work without pay, the Survey applies a test of duration of absence from work. Therefore, a respondent who had been away from work for four weeks or more without pay, is regarded as not employed.

By contrast, the Census does not apply tests of duration for absence from work, and as a result, all persons away from work are most likely to be classified as employed. This of course depends on how the respondent has completed the Census form. As a consequence, a proportion of Census respondents who would be regarded as employed by the Census would be regarded as unemployed or not in the labour force by the Labour Force Survey. As there is no clear way of identifying the Industry of persons classified as employed by the Census but unemployed or not in the labour force by the Survey, it is not possible to remove this population from Census data.

For further information on the Census and the Labour Force Survey, see *Labour Statistics: Concepts, Sources and Methods, 2001* (cat no. 6102.0).

To facilitate reconciliation, the scopes of the 2001 Census and the August 2001 Labour Force Survey were reduced, as far as possible, to a common population. Table 15 below shows the adjustments made to the Labour Force Survey benchmarks and to Census data for Industry.

TABLE 15: ADJUSTMENTS MADE TO AUGUST 2001 LABOUR FORCE SURVEY (LFS) BENCHMARKS AND 2001 CENSUS TO DERIVE A COMMON POPULATION FOR INDUSTRY DATA

Population group	*Deducted from LFS*	*Deducted from Census counts*
Other territories (a)		1,145
Defence force personnel		61,139
Not enumerated in the Census (the Undercount)	289,777	
Residents temporarily overseas	302,323	
Not stated for industry		144,613

(a) Includes Christmas Island, Cocos (Keeling) Islands, and the Jervis Bay Territory.

7.2 Results of Data Reconciliation

The following analyses are based on the 2001 Census and the August 2001 Labour Force Survey. Comparisons by Industry division and age groups, and comparisons by Industry division and states and territories are presented below.

The Census used an additional category, 'Non-classifiable Economic Units' when Industry responses could not be allocated ANZSIC codes. The interviewer-based Labour Force Survey did not require such a category. Therefore, 47,880 Census responses were not distributed to Industry divisions and contributed to the differences between the two collections.

Adjusted August 2001 Labour Force Survey figures for total employed persons were 3.1 per cent (or an estimated 248,777 persons) higher than the figures for the 2001 Census.

7.2.1 Comparison of Industry Divisions by Age using Census Counts as a Proportion of Labour Force Estimates

Table 16 below presents Census Industry by age counts as a proportion of the Labour Force Survey estimates. Tables A1 and A2 in Appendix 2 show the adjusted figures used to derive these proportions. The categories in the Census and in the Labour Force Survey were standardised to reflect the same total population.

TABLE 16: INDUSTRY DIVISION BY AGE, 2001 CENSUS AS A PROPORTION OF AUGUST 2001 LABOUR FORCE SURVEY ESTIMATES

	Age Group						
Industry Division	*15-19*	*20-24*	*25-34*	*35-44*	*45-54*	*55 and over*	*Total*
Agriculture, Forestry and Fishing	0.63	0.66	0.77	0.79	0.89	0.87	0.81
Mining	0.98	1.22	0.97	1.09	0.92	1.65	1.04
Manufacturing	1.11	0.95	0.95	1.04	1.02	1.09	1.01
Electricity, Gas and Water Supply	2.67	0.67	0.83	1.02	0.90	1.11	0.93
Construction	0.73	0.83	0.88	0.93	0.90	1.10	0.90
Wholesale Trade	1.44	1.10	1.15	1.08	1.14	1.22	1.14
Retail Trade	0.87	0.87	0.95	1.07	1.01	1.17	0.96
Accommodation, Cafes and Restaurants	0.89	1.08	0.92	0.97	1.13	1.05	0.99
Transport and Storage	0.69	0.81	0.91	0.88	0.94	1.09	0.92
Communication Services	0.54	1.19	1.02	0.93	0.90	0.91	0.95
Finance and Insurance	1.22	0.85	0.99	0.90	1.00	1.29	0.97
Property and Business Services	0.96	0.83	0.97	0.99	0.99	0.96	0.96
Government Administration and Defence	0.80	0.77	0.96	0.81	0.92	0.85	0.87
Education	0.66	0.83	0.97	1.01	1.04	1.00	0.99
Health and Community Services	0.85	0.87	0.95	1.02	1.04	1.03	0.99
Cultural and Recreational Services	0.93	0.98	0.99	1.21	0.98	1.02	1.03
Personal and Other Services	0.67	0.92	0.80	0.97	0.94	1.05	0.89
Non-classifiable Economic Units	. .	. .	. .	. .	. .	. .	. .
Total	**0.88**	**0.90**	**0.95**	**1.03**	**1.00**	**1.04**	**0.97**

. . Not applicable.

Table 16 above shows that the greatest difference appeared in the lowest age group where Census totals for 15-19 year olds were 88 per cent of the totals for the Labour Force Survey. This is the same result as was obtained in the 1996 reconciliation exercise.

The Industry division Agriculture, Forestry and Fishing recorded the largest proportional difference between the Census and the Labour Force Survey figures. Overall there were 19 per cent fewer respondents in this category for the Census than for the Labour Force Survey.

The second highest proportional difference was for the Industry division, Wholesale Trade where the Census recorded 14 per cent more respondents than the Labour Force Survey.

Within cross-categories 'Industry by age', Labour Force estimates exceeded Census counts by the largest proportions for Agriculture, Forestry and Fishing for 15-19 year olds (by 37 per cent), Transport and Storage for 15-19 year olds (by 31 per cent), Communication for 15-19 year olds (by 46 per cent), Education for 15-19 year olds (by 34 per cent), Personal and Other Services for 15-19 year olds (by 33 per cent), Agriculture, Forestry and Fishing for 20-24 year olds (by 34 per cent), and Electricity, Gas and Water Supply for 20-24 year olds (by 33 per cent).

Census counts exceeded Labour Force estimates by the largest proportions for Electricity, Gas and Water Supply for 15-19 year olds (by 167 per cent), Wholesale Trade for 15-19 year olds (by 44 per cent), Mining for 55 year olds and over (by 65 per cent) and Finance and Insurance for 55 year olds and over (by 29 per cent).

It should be noted that many of these cross-categories (particularly for younger age categories) were represented by small groups which exaggerate the proportional differences. Refer to *Appendix 2 (Tables A1 and A2)* for counts/estimates.

7.2.2 Comparison of Industry Divisions by State and Territory Using Census Counts as a Proportion of Labour Force Estimates

Table 17 presents Census Industry by state and territory counts as a proportion of the Labour Force estimates. Tables A3 and A4 in Appendix 2 show the adjusted figures by state and territory used to derive these proportions. The categories in the Census and in the Labour Force Survey were standardised to reflect the same total population in each state or territory.

TABLE 17: INDUSTRY DIVISION BY STATE AND TERRITORY, 2001 CENSUS AS A PROPORTION OF AUGUST 2001 LABOUR FORCE SURVEY ESTIMATES

	State/ Territory							
Industry Division	*NSW*	*Vic*	*Qld*	*SA*	*WA*	*Tas*	*NT*	*ACT*
Agriculture, Forestry and Fishing	0.78	0.87	0.80	0.87	0.79	0.72	0.68	0.65
Mining	0.82	1.28	1.07	0.92	1.12	0.99	2.87	. .
Manufacturing	1.01	0.96	1.05	1.10	1.03	1.06	1.15	0.99
Electricity, Gas and Water Supply	0.98	0.69	1.03	0.90	1.08	1.15	2.79	1.38
Construction	0.92	0.94	0.85	0.93	0.82	0.92	1.14	1.01
Wholesale Trade	1.11	1.23	1.09	1.12	1.15	1.24	1.05	0.81
Retail Trade	0.96	0.93	1.00	1.01	0.95	1.00	0.93	1.03
Accommodation, Cafes and Restaurants	1.00	1.09	0.95	1.04	0.79	1.20	0.93	1.48
Transport and Storage	0.93	0.83	0.96	0.98	0.91	1.14	0.97	1.16
Communication Services	0.91	0.93	0.97	1.05	1.01	1.12	0.64	1.39
Finance and Insurance	0.92	0.95	1.10	0.97	1.06	1.26	1.20	1.07
Property and Business Services	0.95	1.01	0.93	0.94	0.90	1.05	1.05	0.95
Government Administration and Defence	0.92	0.72	0.93	0.99	1.03	0.89	0.55	0.92
Education	0.97	0.97	1.02	0.95	0.99	1.22	0.89	1.22
Health and Community Services	0.94	1.00	1.06	1.12	0.94	0.98	1.04	0.88
Cultural and Recreational Services	1.08	1.00	1.07	0.93	0.97	0.92	1.37	0.93
Personal and Other Services	0.90	0.89	0.91	0.97	0.84	0.84	0.94	0.78
Non-classifiable Economic Units	. .	. .	. .	. .	. .	. .	. .	. .
Total	**0.96**	**0.97**	**0.98**	**1.01**	**0.95**	**1.02**	**0.93**	**1.00**

. . Not applicable.

The major proportional differences across the states and territories between the two collections occurred primarily in the Northern Territory and the Australian Capital Territory, with notable differences in Victoria.

In the Northern Territory, Labour Force estimates significantly exceeded Census counts in Agriculture, Forestry and Fishing (by 32 per cent), in Communication Services (by 36 per cent), and in Government Administration and Defence (by 45 per cent), whereas Census counts significantly exceeded Labour Force estimates in Mining (by 187 per cent), in Electricity, Gas and Water Supply (by 179 per cent), and in Cultural and Recreational Services (by 37 per cent).

In the Australian Capital Territory, Labour Force estimates significantly exceeded Census counts in Agriculture, Forestry and Fishing (by 35 per cent) whereas Census counts significantly exceeded labour Force estimates in Electricity, Gas and Water Supply (by 38 per cent), in Accommodation, Cafes and Restaurants (by 48 per cent), and Communication Services (by 39 per cent)

These differences in the Northern Territory and the Australian Capital Territory may reflect sampling variability in the smaller population areas in the Labour Force Survey.

The large proportional difference of 28 per cent for Mining in Victoria probably reflects the small population in this category (4,472 persons in the Census and 3,481 in the Labour Force estimates). Also in Victoria, a significantly higher number of persons were identified as being employed in the Electricity, Gas and Water Supply Industry in the Labour Force estimates (18,732) than in the Census (12,916), a proportional difference of 31 per cent.

8. CONCLUSIONS

This paper has examined the quality of Industry data from the 2001 Census. The main conclusions are:

- 55.1 per cent of 2001 Census Industry data were coded automatically whilst the remaining 44.9 per cent were coded by the Computer Assisted Coding (CAC) and Query Resolution (QR) processes. Users of Industry data should be aware that the two coding procedures yielded different data distributions.

- The non-response rate for Industry decreased only marginally from 2.0 per cent in 1996 to 1.7 per cent in 2001 with the changes in form design helping to maintain a favourable rate.

- The Industry division Transport and Storage contained the highest level of undefined coding with only 79.5 per cent of the responses coded to the ANZSIC class level. Manufacturing division recorded the next highest level with only 83.0 per cent of responses coded to the most defined level. The introduction of Intelligent Character Recognition (ICR) processing doesn't appear to have brought about any improvement to the level of responses coded to the most detailed ANZSIC level for these Industry divisions when compared to 1996.

- The improved level in 2001 of responses coded to the most detailed ANZSIC level in Agriculture, Forestry and Fishing (up 26.1 percentage points) can be attributed to the work by classifications staff on the Agriculture and Mining areas of the coding index. In 1996 Agriculture, Forestry and Fishing contained the highest level of undefined coding which was attributed to the reliance by coders on the often inadequate description by respondents (e.g. 'farmer').

- Mining also improved, with the level of defined coding increasing by 21.1 percentage points. Improved specification of the mined product (e.g. coal mining) and additional index entries, contributed to the fall in the proportion of responses coded to the 1-digit level.

- Discrepancy analyses showed that for some codes within the Construction Trade Services subdivision, the AC process had allocated codes for Carpentry Services, Bricklaying Services and Plastering and Ceiling Services when the correct code should have been for House Construction. In 1996, coders had difficulty determining whether a Construction response was a General Construction response (incorporating Building Construction and Non-building Construction) or a more specialised Construction Trade Service response (incorporating Building Structure Services and Installation Trade Services).

- Computer Assisted Coding (CAC) discrepancies were mostly due to coders obtaining a code when a query should have been raised.

- Data reconciliation between the 2001 Census and the August 2001 Labour Force Survey showed that the differences in the counts/estimates between the two collections were statistically significant, as was the case in 1996.

- For the 2001 Census, only marginal improvement in the quality of the responses can be attributed to the use of the two-part Industry question which was expected to better identify the activity and products of the employer's business, than the nature of the business. However, the use of Automatic Coding (AC) and the 'structured' Industry coding index for 2001 has reduced inconsistencies in coding which can be introduced by coders through varying levels of knowledge and different attitudes.

9. RECOMMENDATIONS

- The editing/coding strategy needs to be well tested, and finalised before the 2006 Census. The strategy should not be changed or augmented part the way through topic coding, unless all previously edited data are reprocessed. Further improvements to the index set-up and consistent paths for the two processes, AC and CAC, are required before consistent codes across them can be achieved.

- Conceptually the 2001 Census had four Industry-based questions, while the 1996 Census had just three questions. To maximise coding matching in 2006, using either a business-based index (Inteframe) or an output /activity based index ('Structured' Industry Coder), all four questions could be retained. However, considering the limited use made of the mark-box question, it should either be dropped, or strengthened by using a full list of Industry divisions as was recommended in the September 1998 tests.

- The addition of postcode as a business address field could give the coder greater discretion in cases where an exact match is not possible but other available information suggests that a match is likely. A complete list of Inteframe units, groomed to allow for automatic repair issues likely to arise, and the requirement to code to location level, would further assist in achieving a higher match rate.

- QM needs information about the 'severity' of the discrepancies to better measure data quality in terms of a valid data outcome, as well as to respond to the procedural issue of whether to raise a query or not.

- The finding, as a result of a DQI that one third of respondents employed in an industry other than the five listed on the form, ignored marking the 'Other - please specify' box, instead going direct to the write-in field to answer the question, supports the removal of a mark-box for all 'Other - please specify' options on future Census forms. The high level of completion suggests that it may not be necessary to have a two-stage process (i.e. an 'Other' mark-box, plus write-in boxes) to elicit such information. The omission of a mark-box for 'Other - please specify' will also eliminate the occurrence of people marking 'Other' but not supplying further business information. It is recommended that the requirement for the 'Other- please specify' for this, and similar mark-box questions on the form, is tested before the next Census.

APPENDIX 1: Example of Australian and New Zealand Standard Industrial Classification (ANZSIC) - Division, Subdivision, Group and Class

E CONSTRUCTION

41 General Construction

410 General Construction, undefined
- 4100 General Construction, undefined
- 4110 Building Construction , undefined
- 4111 House Construction
- 4112 Residential Building Construction, undefined
- 4113 Non-Residential Building Construction

412 Non-Building Construction
- 4120 Non-Building Construction, undefined
- 4121 Road & Bridge Construction
- 4122 Non-Building Construction, not elsewhere classified

42 Construction Trade Services

420 Construction Trade Services, undefined
- 4200 Construction Trade Services, undefined

421 Site Preparation Services
- 4210 Site Preparation Services

422 Building Structure Services
- 4220 Building Structure Services, undefined
- 4221 Concreting Services
- 4222 Bricklaying Services
- 4223 Roofing Services
- 4224 Structural Steel Erection Services

423 Installation Trade Services
- 4230 Installation Trade Services, undefined
- 4231 Plumbing Services
- 4232 Electrical Services
- 4233 Air Conditioning and Heating Services
- 4234 Fire and Security System Services

424 Building Completion Services
- 4240 Building Completion Services, undefined
- 4241 Plastering and Ceiling Services
- 4242 Carpentry Services
- 4243 Tiling and Carpeting Services
- 4244 Painting and Decorating Services
- 4245 Glazing Services

425 Other Construction Services
- 4250 Other Construction Services, undefined
- 4251 Landscaping Services
- 4259 Construction Services, not elsewhere classified

E0 Construction, undefined

E00 Construction, undefined
- E000 Construction, undefined

APPENDIX 2: Reconciliation between 2001 Census and August 2001 Labour Force Survey - adjusted data tables

TABLE A1: ADJUSTED FIGURES FOR INDUSTRY DIVISION BY AGE, 2001 CENSUS

	15-19	*20-24*	*25-34*	*35-44*	*45-54*	*55 and over*	*Total*
Agriculture, Forestry and Fishing	13,205	21,361	56,240	74,139	75,543	90,288	330,776
Mining	990	4,567	20,339	24,091	18,725	6,344	75,056
Manufacturing	37,916	88,931	251,818	282,961	231,967	116,575	1,010,168
Electricity, Gas and Water Supply	887	3,538	13,563	18,292	17,762	6,621	60,663
Construction	26,729	55,307	141,043	151,555	120,732	63,132	558,498
Wholesale Trade	18,173	42,944	113,161	116,320	95,072	51,448	437,118
Retail Trade	257,826	176,547	242,894	231,926	199,440	102,631	1,211,264
Accommodation, Cafes and Restaurants	57,854	76,839	93,724	80,129	67,558	34,427	410,531
Transport and Storage	6,497	24,085	83,652	100,123	90,590	50,869	355,816
Communication Services	3,017	13,481	41,993	42,391	35,724	11,871	148,477
Finance and Insurance	6,581	35,746	104,016	81,446	60,662	23,936	312,387
Property and Business Services	30,232	97,787	246,785	232,600	203,845	109,027	920,276
Government Administration and Defence	5,165	19,465	70,643	88,621	87,726	35,873	307,493
Education	8,299	35,704	112,623	168,082	195,284	75,281	595,273
Health and Community Services	18,334	58,673	169,660	232,066	225,197	102,187	806,117
Cultural and Recreational Services	18,304	27,569	53,508	47,751	35,433	19,872	202,437
Personal and Other Services	17,541	29,182	77,929	78,787	64,680	32,515	300,634
Non-classifiable Economic Units	2,695	4,785	10,733	11,905	10,547	7,215	47,880
Total	**530,245**	**816,511**	**1,904,324**	**2,063,185**	**1,836,487**	**940,112**	**8,090,864**

TABLE A2: ADJUSTED FIGURES FOR INDUSTRY DIVISION BY AGE, AUGUST 2001 LABOUR FORCE SURVEY

	15-19	*20-24*	*25-34*	*35-44*	*45-54*	*55 and over*	*Total*
Agriculture, Forestry and Fishing	20,797	32,234	73,496	93,644	84,845	103,695	408,711
Mining	1,008	3,733	20,936	22,073	20,322	3,847	71,919
Manufacturing	34,055	93,402	265,439	271,417	226,334	107,137	997,786
Electricity, Gas and Water Supply	332	5,256	16,279	17,999	19,688	5,960	65,504
Construction	36,623	66,974	161,191	163,542	133,884	57,371	619,584
Wholesale Trade	12,644	38,918	98,430	107,833	83,642	42,164	383,632
Retail Trade	296,046	203,344	254,877	217,120	197,415	87,871	1,256,673
Accommodation, Cafes and Restaurants	64,982	71,409	101,709	82,509	59,608	32,667	412,884
Transport and Storage	9,476	29,631	91,899	113,197	96,035	46,656	386,894
Communication Services	5,573	11,318	41,122	45,828	39,909	13,085	156,835
Finance and Insurance	5,376	41,941	105,011	90,587	60,673	18,565	322,154
Property and Business Services	31,352	117,380	254,415	235,817	205,717	113,909	958,950
Government Administration and Defence	6,448	25,145	73,717	108,761	95,249	42,249	351,570
Education	12,561	43,075	115,631	167,232	187,690	75,466	601,657
Health and Community Services	21,504	67,207	178,894	227,897	216,809	99,431	811,743
Cultural and Recreational Services	19,670	28,059	54,058	39,470	36,078	19,575	196,910
Personal and Other Services	26,228	31,841	97,144	81,240	69,099	31,045	336,596
Non-classifiable Economic Units	. .	. .	. .	. .	. .	. .	. .
Total	**604,678**	**910,867**	**2,004,239**	. .	**1,832,997**	**900,694**	**8,339,641**

. . Not applicable.

TABLE A3: ADJUSTED FIGURES FOR INDUSTRY DIVISION BY STATE AND TERRITORY, 2001 CENSUS

	NSW	*Vic*	*Qld*	*SA*	*WA*	*Tas*	*NT*	*ACT*
Agriculture, Forestry and Fishing	92,358	72,639	76,532	36,867	36,674	12,261	2,788	653
Mining	14,823	4,472	19,286	3,864	28,771	1,550	2,215	72
Manufacturing	316,113	318,218	167,380	93,428	84,281	21,125	4,059	5,562
Electricity, Gas and Water Supply	20,389	12,916	12,359	4,640	6,878	1,787	781	913
Construction	189,740	136,454	111,209	36,463	61,961	9,326	5,594	7,732
Wholesale Trade	152,790	115,909	79,718	31,561	42,305	8,402	3,274	3,153
Retail Trade	390,914	307,419	239,615	92,549	123,049	27,354	10,729	19,633
Accommodation Cafes and Restaurants	141,927	90,302	88,381	28,704	38,321	9,458	5,440	7,996
Transport and Storage	125,752	79,010	77,587	24,005	32,630	7,899	4,762	4,166
Communication Services	54,958	41,826	23,016	10,334	12,115	2,781	1,020	2,420
Finance and Insurance	131,955	81,986	44,562	19,935	24,121	4,443	1,541	3,840
Property and Business Services	334,299	237,123	153,864	59,374	90,141	14,113	7,673	23,684
Government Administration and Defence	87,568	52,967	61,942	22,560	32,702	9,332	9,814	30,600
Education	187,168	147,473	118,896	44,933	60,318	15,040	7,179	14,276
Health and Community Services	258,522	202,226	151,029	72,441	79,276	21,261	7,976	13,372
Cultural and Recreational Services	67,595	53,251	37,341	13,238	18,220	4,310	2,655	5,830
Personal and Other Services	98,321	69,531	57,662	24,433	33,104	6,888	4,305	6,388
Non-classifiable Economic Units	14,884	11,681	7,452	3,088	7,093	1,508	865	1,309
Total	**2,680,076**	**2,035,403**	**1,527, 831**	**622,417**	**811,960**	**178,838**	**82,670**	**151,599**

TABLE A4: ADJUSTED FIGURES FOR INDUSTRY DIVISION BY STATE AND TERRITORY, AUGUST 2001 LABOUR FORCE SURVEY

	NSW	*Vic*	*Qld*	*SA*	*WA*	*Tas*	*NT*	*ACT*
Agriculture, Forestry and Fishing	118,155	83,438	96,126	42,263	46,687	16,938	4,101	1,005
Mining	18,131	3,481	18,040	4,218	25,708	1,570	771	0
Manufacturing	311,570	330,014	159,940	85,081	82,079	19,939	3,535	5,628
Electricity, Gas and Water Supply	20,761	18,732	11,985	5,151	6,375	1,559	280	660
Construction	206,372	144,516	131,072	39,351	75,641	10,114	4,886	7,632
Wholesale Trade	137,404	94,351	73,257	28,174	36,656	6,796	3,109	3,886
Retail Trade	409,052	329,126	239,075	91,524	129,860	27,470	11,527	19,041
Accommodation, Cafes and Restaurants	141,251	83,151	93,304	27,572	48,468	7,868	5,849	5,420
Transport and Storage	135,556	94,973	80,541	24,565	35,847	6,937	4,895	3,580
Communication Services	60,531	44,996	23,684	9,821	11,983	2,481	1,604	1,735
Finance and Insurance	143,795	85,896	40,587	20,557	22,753	3,526	1,284	3,574
Property and Business Services	350,580	233,831	165,298	63,017	100,244	13,504	7,287	24,831
Government Administration and Defence	95,291	73,682	66,479	22,864	31,719	10,451	17,730	33,353
Education	193,704	151,493	116,040	47,379	61,007	12,298	8,043	11,693
Health and Community Services	274,161	202,194	142,008	64,693	84,221	21,696	7,646	15,124
Cultural and Recreational Services	62,459	53,428	35,055	14,307	18,761	4,709	1,936	6,256
Personal and Other Services	109,736	77,791	63,380	25,177	39,480	8,235	4,561	8,239
Non-classifiable Economic Units	. .	. .	. .	. .	. .	. .	. .	. .
Total	**2,788,690**	**2,105,093**	**1,555,869**	**615,713**	**857,489**	**176,090**	**89,042**	**151,655**

. . Not applicable.

GLOSSARY

Australian Bureau of Statistics (ABS)	Australia's official statistical agency.
Australian and New Zealand Standard Industrial Classification (ANZSIC)	A classification, first issued in 1993, developed for use in Australia and New Zealand for the production and analysis of industry statistics. For more information refer to Appendix 1 or the *Australian and New Zealand Standard Industrial Classification (ANZSIC) 1993* (cat. no.1292.0).
Australian Standard Geographical Classification (ASGC)	A geographic classification system for identifying states, parts of states and smaller areas, in a uniform manner.
Automatic Coding (AC)	A system used to automatically allocate codes to the data stored by the Intelligent Character Recognition (ICR) system following the scanning of the Census forms.
Collection District (CD)	The smallest geographical area covered by the Census, as defined by the ASGC. It usually relates to an area allocated to a Census collector in which they deliver and collect Census forms.
Community Development Employment Program (CDEP)	An employment program available to Indigenous people.
Computer Assisted Coding (CAC)	A system which helps coders to classify written responses on Census forms using a structured coding index.
Data Capture (DC)	The process that ensures that marks on the Census form (mark-box or writing) are reproduced on an image. DC registers and codes mark-box responses.
Data Processing Centre (DPC)	The centralised facility for processing the 2001 Census forms located in Ultimo, NSW.
Data Quality Investigation (DQI)	A DQI team operated at the DPC, conducting additional coding exercises to uncover data quality issues.
Discrepancy Rate	The rate at which QM and subsequent adjudication coding differed from that of an individual coder or system coding. It is expressed as a percentage and is regarded as the error rate within final data.
Intelligent Character Recognition (ICR)	A system which scans Census forms, reads the hand-printed data, verifies and corrects the data read from the form, and stores the form image and data for additional processing.
Labour Force Survey (LFS)	An ABS interviewer-based survey conducted monthly. The purpose of the LFS is to provide timely information on the labour market activity of the civilian population of Australia aged 15 years and over. It is the official source for the labour force participation and unemployment rates.

Management Information System (MIS)	A DPC-based system that accumulated and produced statistics on the progress and quality of the processing operation.
Mark-boxes	Boxes that invite the respondent to place a dash on one of a possible series of selection boxes on the Census form. The ICR system then identified marked boxes during DC.
Non-Classifiable Economic Units	When industry responses can not be allocated ANZSIC codes because they contain insufficient information, the Census uses an additional category, 'Non-classifiable Economic Units'. The interviewer-based collections, such as the LFS, do not require such a category interviewers are able to obtain codeable responses. This factor contributes to Industry-related differences between the Census and other ABS collections.
Other territories	Since the 1996 Census, Christmas Island, Cocos (Keeling) Islands, and the Jervis Bay Territory (previously linked to the Australian Capital Territory for statistical purposes) comprise a pseudo 'ninth state/territory' of Australia.
Quality Management (QM)	The process of regular review of a percentage of coding work. Also a term for broader DPC-wide ongoing reviews.
Query Resolution (QR)	A specialist group with access to additional resource material who resolved difficult coding issues.
Repair	Comprises a two stage manual process after initial scanning of the forms. First, a high speed repair method displays individual characters (carpets) for confirmation and unknown/unsure characters (triplets) in sets of three for key entry. A second stage involves fields still requiring repair being displayed for key entry repair.
Second Release Processing (SRP)	Responses to the more complex Census topics, such as Industry, were processed within this second phase.
Self-enumeration	Is the term used to describe the way Census data are collected. Census forms are generally completed by householders (or individuals in non-private dwellings) rather than by interviewers, although interviewers are available in some areas, such as Indigenous communities.
System Created Record (SCR)	Is a record created during Census processing for a person for whom a Census form has not been received but where the collector believed the dwelling was occupied on Census night. These records have values imputed for age, sex, marital status and usual residence only. Values for other variables are set to Not Stated or Not Applicable depending on the imputed value for age.
Write-in Response Box	A response box on the Census form requiring a written text or numeric response, generally coded using ICR and then AC.

Note

For more information about the terms, definitions and descriptions of categories in this paper refer to the *2001 Census Dictionary*, (cat. no. 2901.0).

REFERENCE LIST

Australian Bureau of Statistics/New Zealand Department of Statistics, *Australian and New Zealand Standard Industrial Classification (1993)*, (cat. no. 1292.0), Commonwealth of Australia, Canberra

Australian Bureau of Statistics, *2001 Census Dictionary*, (cat. no. 2901.0)

Australian Bureau of Statistics, *Labour Force, Australia, August 2001*, (cat. no. 6203.0)

Australian Bureau of Statistics, *Labour Statistics: Concepts, Sources and Methods, 2001* (cat no. 6102.0).

Census Working Paper 00/3*: 1996 Census Data Quality: Industry*

CENSUS PAPERS

2001 Census Papers:

03/09	*2001 Census: Level, Main Field and Year of Completion of Highest Non-School Qualification*
03/08	*2001 Census: Industry*
03/06	*2001 Census: Occupation*
03/05	*2001 Census: Labour Force Status*
03/04	*2001 Census: Income*
03/03	*2001 Census: Computer and Internet Use*
03/02	*2001 Census: Housing*
03/01b	*2001 Census: Ancestry - Detailed Paper*
03/01a	*2001 Census: Ancestry - First and Second Generation Australians*
02/03	*2001 Census: Form Design Testing*
02/02	*Report on Testing of Disability Questions for Inclusion in the 2001 Census*
02/01	*2001 Census: Digital Geography Technical Information Paper*

1996 Census Working Papers:

00/4	*1996 Census Data Quality: Income*
00/3	*1996 Census Data Quality: Industry*
00/2	*1996 Census Data Quality: Qualification Level and Field of Study*
00/1	*1996 Census Data Quality: Journey to Work*
99/6	*1996 Census Data Quality: Occupation*
99/4	*1996 Census: Review of Enumeration of Indigenous Peoples in the 1996 Census*
99/3	*1996 Census Data Quality: Housing*
99/2	*1996 Census: Labour Force Status*
99/1	*1996 Census: Industry Data Comparison*
97/1	*1996 Census: Homeless Enumeration Strategy*
96/3	*1996 Census of Population and Housing: Digital Geography Technical Information Paper*
96/2	*1996 Census Form Design Testing Program*

A range of 1991 Census Working Papers, from 93/1 to 96/1 are also available.

These Papers can be accessed on the ABS web site at <http://www.abs.gov.au>. From the ABS home page, select **Census -> (Census Information) Fact Sheets and Census Papers -> (Fact Sheets and Information Papers) Census Papers.**

If you have further data quality queries, please contact the Assistant Director, Census Evaluation by telephone: (02) 6252 5611 or email: <joanne.healey@abs.gov.au>.