1351.0.55.026

Australian Bureau of Statistics

Research Paper

Assessing the Likely Quality of the Statistical Longitudinal Census Dataset

www.abs.gov.au

New Issue

Research Paper

Assessing the Likely Quality of the Statistical Longitudinal Census Dataset

Glenys Bishop

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 19 AUG 2009

ABS Catalogue no. 1351.0.55.026

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Jonathon Khoo, Analytical Services Branch on Canberra (02) 6252 5506 or email <analytical.services@abs.gov.au>.

CONTENTS

ASSESSING THE LIKELY QUALITY OF THE STATISTICAL LONGITUDINAL CENSUS DATASET

Glenys Bishop
Analytical Services

ABSTRACT

As part of the Census Data Enhancement project, the Australian Bureau of Statistics has conducted a quality study that simulates the formation of the Statistical Longitudinal Census Dataset (SLCD). This simulation has been carried out by linking 2006 Census and 2006 Census Dress Rehearsal data. The linking was carried out both with and without the use of name and address, the former acting as a benchmark for the latter. Linking without name and address is the method that is likely to be used for the planned linking of the 2006 and 2011 Censuses forming the first two waves of the SLCD. This paper describes a variety of methods that have been used to examine the quality of the data linked without name and address and extends those findings to predict the quality that can be expected when the first two waves of the SLCD are linked.

1. INTRODUCTION

In April 2005, the Australian Bureau of Statistics (ABS) released a discussion paper (ABS, 2005b) seeking public comment on a proposal for the Census Data Enhancement project. Following a number of submissions and a Privacy Impact Assessment, the Statistician released a Statement of Intention (ABS, 2005a) describing how this project would be conducted.

The key feature of the Census Data Enhancement project is the formation of a Statistical Longitudinal Census Dataset (SLCD) which entails choosing a random sample of 5% of persons in the 2006 Census of Population and Housing and endeavouring to bring together those persons' 2006 records with their records from the 2011 and subsequent Censuses. At each Census, the SLCD will be augmented with a 5% sample of children, who have been born, and immigrants, who have arrived, since the previous Census. There will also be some provision for topping up the sample to maintain a dataset that is consistently 5% of the Census population. As a result of the extensive public consultation, and in line with ABS policy, the links between records from successive Censuses will have to be made without using name and address.

In the Statement of Intention, the Statistician also described two other key features of the Census Data Enhancement project. One feature is that approved *statistical studies* may be conducted by linking the SLCD with other specified datasets. The other feature is that *quality studies* can be conducted by linking the complete Census dataset with other specified datasets. If this linking is performed during the Census processing period, when name and address are available, then name and address can be used in the linking process. It should be noted that all names and addresses are destroyed at the end of the Census processing period.

An information paper (ABS, 2006a) contains further details of the quality studies and a statistical study that were planned for the 2006 Census.

This paper reports on the findings of one of those quality studies.

2. THE SIMULATED FORMATION OF THE SLCD QUALITY STUDY

One quality study that was conducted during the 2006 Census processing period was the Simulated Formation of the SLCD. This quality study aimed to assess the feasibility of linking a 5% sample of the 2006 Census with the 2011 and subsequent Censuses to form the Statistical Longitudinal Census Dataset, without using name and address, and to make defensible statements about the likely quality of the linkage. The simulation has been performed by linking the records of persons in the Census Dress Rehearsal, which was conducted one year prior to the 2006 Census, to the records of the same persons as far as possible in the 2006 Census.

This quality study has been used not only as a vehicle to develop linkage methods and associated tools but also to research and implement methods for analysis of linked data. Early ABS thinking about linkage methods is described in Conn and Bishop (2006) and an early summary of methods considered by the ABS to assess quality of linked data is described in Bishop and Khoo (2007).

3. THE DATA

A full description of the data and the linkage process is given by Solon and Bishop (2009). However, some of the key features of the Census data are mentioned here to provide context for the rest of the discussion.

The Census Dress Rehearsal conducted one year before the 2006 Census contained exactly the same questions and formats for self-completion. Several different types of forms were used in both. Household forms allowed for data to be collected for up to six members of a household, personal forms accommodated only one person, electronic forms could be completed by either households or individuals and special Indigenous forms were used by interviewers to collect information from Indigenous persons in remote areas. Other collection methods used in the Census, such as administrative forms, were not used in the Census Dress Rehearsal.

The Census Dress Rehearsal collected data from persons who were in selected collection districts on the night of 9 August 2005. These collection districts were in parts of Sydney, Wagga Wagga and Junee in New South Wales and in parts of Adelaide and Murray Bridge in South Australia. Data were also collected from three remote communities in Western Australia and the Northern Territory. Details are contained in an information paper (ABS, 2006b).

The Census file used in this study contained records for 19,050,146 persons. Overseas visitors and imputed persons were excluded from the file. The latter are people known to exist but for whom no Census form was returned and so a statistical method was used to impute their demographic information. See the ABS (2008b) publication.

In both collections, respondents were asked to provide the address where they were on Census night, on the front of the form, their usual address if that was different, their usual address one year ago and their usual address five years ago. Full descriptions of these addresses are available in Solon and Bishop (2009).

A new building block of statistical and administrative geography that was introduced with the 2006 Census was the mesh block. These are micro-level geographical units for statistics and altogether there are in excess of 300,000 mesh blocks covering the whole of Australia. A mesh block may contain a residential area, an administrative area such as Parliament House or a geographic feature such as a national park. Typically, a residential mesh block contains between 30 and 60 dwellings. A street address can be coded to the appropriate mesh block but, since many such addresses will occur in any given mesh block, it is not possible to convert a mesh block to an address. Mesh blocks serve as a very useful geographical indicator and have been used in linking Census Dress Rehearsal and Census data. Further information is available in two information papers (ABS, 2007a and 2008a).

4. FORMING LINKED DATASETS

The linking methodology used to link the Census Dress Rehearsal and the 2006 Census data, either with or without name and address, was probabilistic linking. The method links records from two files using several variables common to both files. A key feature of this methodology is its ability to handle a variety of linking variables and record comparison methods to produce a single numerical measure of how well two particular records match. This allows ranking of all possible links and optimal assignment of the link or non-link status.

The first linked dataset was formed using name, address, mesh block and other variables as linking variables; it was produced through exhaustive linking passes and by extensive clerical review. This dataset was created with the intention of being used as a benchmark for other linked datasets and was termed the Gold Standard.

The Silver Standard was formed by using hash value, mesh block and other variables as linking variables. A one-way encoding algorithm was developed to convert a combination of first name and surname to a number between 1 and 12,005, termed the hash value. There was a minimum of 1,500 distinct names corresponding to any given hash value. While the algorithm will always convert a particular combination of first name and surname to the same hash value, it is not possible to derive names from a given hash value. Hash values were deleted along with names and addresses at the end of the Census processing period.

The Bronze Standard was formed through use of mesh block and other variables as linking variables. This mimics the method to be used in forming the SLCD.

During the linking process, pairs of records, one from each dataset to be linked, are assigned a weight derived from the level of agreement in the linking variables. A cut-off weight is set and those pairs above the cut-off are declared links. For each of the Silver and Bronze Standards, four linked datasets were produced, corresponding to Very Low, Low, Medium and High cut-offs.

The Very Low cut-off linked datasets were generated by setting the record-pair weights cut-off at very low points. These datasets contain larger numbers of links compared to the higher cut-off datasets but also contain a larger number of false links. These false links were the result of linking records of people with similar characteristics but not belonging to the same individuals. On the other hand, the High cut-off datasets had the weight cut-offs set high and so the links consisted of a higher proportion of true links but many matches were missed. A match is defined as a pair of records that belong to the same individual.

5. METHODS OF EVALUATION OF THE LINKAGE

There are several ways to evaluate the quality of the linked datasets. For this quality study, the ABS has considered the following:

- the quality of the Gold Standard linked dataset;
- the match-link rate and link accuracy of the different Silver and Bronze Standard linkages compared with the Gold Standard;
- the over- or under-representation of subgroups in the various linked datasets compared with the Gold Standard;
- the effects of this over- or under-representation on some representative analyses and models fitted to linked data;
- methods for modifying the fitted models to account for inexact linkage and disparities in the representation of subgroups of interest;
- how well linking two files collected one year apart can represent linking two files collected five years apart.

These issues are each discussed in turn in Sections 6 to 11 of this paper.

6. THE QUALITY OF THE GOLD STANDARD LINKED DATASET

The Census Dress Rehearsal (CDR) dataset consisted of 78,349 person records after persons who had died between the CDR night and Census night had been removed. Solon and Bishop (2009) have further details. Of the 78,349 records, 8,075 could not be linked to the Census when forming the Gold Standard. Some of these would not have a corresponding Census record. It is important to consider how many records we might reasonably expect to link. Persons on the CDR file might be missing from the Census file for several reasons:

- they are temporarily out of the country on Census night, as described in the ABS (2009) publication;
- they are missed by the Census, thus contributing to Census undercount, as described in the ABS (2007c) publication;
- they do not personally complete a form but it is possible to impute them using information from a reliable source, as described in the ABS (2008b) publication;
- they emigrated from Australia before the Census and details can be found in the ABS (2009) publication; or
- they died before the Census but missed being removed form the CDR file, as described by Solon and Bishop (2009).

To quantify the number of CDR respondents who might have missed being counted in the Census, undercount factors from the Post Enumeration Survey for various demographic cross-tabulations were used. Each CDR record was assigned a probability of being counted in the Census, the probabilities were summed and an expected number of matches was calculated. Using the 'age-by-sex' undercount factors, 5,248 CDR respondents were estimated to have been missed by the Census.

On 2006 Census night, 345,200 people were temporarily out of the country (ABS, 2007b). A pro rata estimate of the number of CDR respondents who were overseas on Census night is 1,301.

The first two reasons for being missed in the Census give an estimated 6,549 missing persons, i.e. 71,800 should be linked. In fact, only 70,274 CDR records were linked to Census records in the Gold Standard, a discrepancy of 1,526.

Solon and Bishop (2009) explore the properties of the 8,075 CDR records that were not linked in the Gold Standard. It is not known which of these records were not linked for reasons described above. However, it is known that just under 500 were missing at least one of their names and, of these, just under 400 were also missing some component of their date of birth, while 128 were also missing mesh block.

About 1,800 were missing all components of date of birth and 1,200 were missing mesh block. It therefore seems likely that approximately 1,500 CDR records were not linked because of poor quality data.

A summary of missing values is shown in table 6.1.

6.1 Missing values from selected linking fields used in the Gold Standard linkage. These numbers apply to the 8,075 Census Dress Rehearsal records that were not linked.

Field	*Number of missing values*
First name	358
Surname	344
First name OR Surname	488
First name AND Surname	214
Mesh block	1,211
First name OR Surname OR Mesh block	1,571
First name AND Surname AND Mesh block	68
(First name OR Surname) AND Mesh block	128
Any component of Date of birth	1,934
All components of Date of birth	1,873
(First name OR Surname) AND any component of Date of birth	378
Street-name OR Suburb	1,160
Street-name AND Suburb	870

7. MATCH-LINK RATE AND LINK ACCURACY

If we consider the Gold Standard links as matches, i.e. each record in a pair relates to the same person, then we can use them as a benchmark for the Bronze and Silver Standard linkages. We are interested in the proportion of links in a given dataset that are matches, and we have termed this the link accuracy. We are also interested in the proportion of matches that are actually linked in the given dataset, and we have termed this the match-link rate. Match-link rate and link accuracy were calculated for each linkage level of the Bronze and Silver Standards by comparing them with the Gold Standard as shown in table 7.1.

Although cells for total numbers of non-links and non-matches are indicated in table 7.1, it is not possible to obtain these numbers when a multiple pass linking methodology is used. Some record-pairs will be compared more than once and some will never be compared.

7.1 Method of calculating Match-link rate and Link accuracy

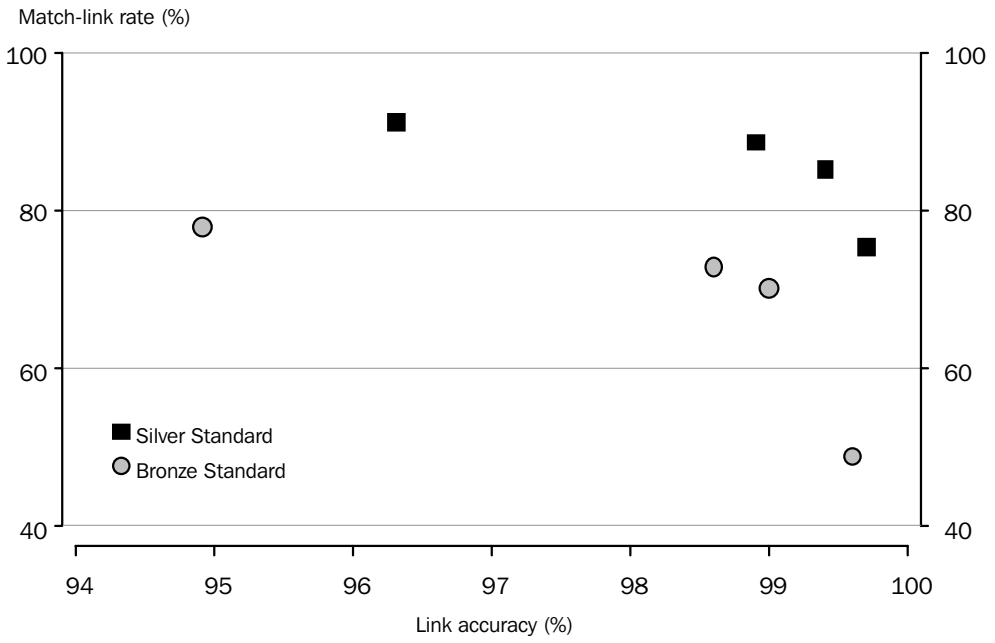
		Match status from Gold Standard		
		Matches	Non-matches	
Link status from Bronze or Silver Standard	Links	(True links)	(Falsely linked)	(Total links)
	Non-links	(Falsely non-linked)	(True non-links)	(Total non-links)
		(Total matches)	(Total non-matches)	

$$\text{Match-link rate} = \frac{\text{True links}}{\text{Total matches}}$$

$$\text{Link accuracy} = \frac{\text{True links}}{\text{Total links}}$$

Results for the High, Medium, Low, and Very Low linked datasets for each of the Bronze and Silver Standards are shown in figure 7.2. As the link accuracy is increased by raising the cut-off, there is a gradual decline in match-link rate until at some point there is a rapid decline. In general, higher cut-offs have fewer links but they are better quality and so have higher link accuracy but lower match-link rates. For a given link accuracy, Silver Standard match-link rates are considerably higher then those for Bronze Standard. The actual numbers of CDR records linked in each dataset are shown in table 7.3.

7.2 Match-link rate and Link accuracy for each Silver and Bronze Standard linkage compared with the Gold Standard linkage



7.3 Number of records linked on each of the datasets, the corresponding proportion of CDR records, Link accuracy and Match-link rate

Dataset	*Number of records linked*	*Proportion of CDR records linked**	*Link accuracy (%)*	*Match-link rate (%)*
Gold Standard	70,274	0.90		
Silver Standard				
High	53,231	0.68	99.7	75.5
Medium	60,367	0.77	99.4	85.4
Low	63,044	0.80	98.9	88.8
Very Low	66,614	0.85	96.3	91.3
Bronze Standard				
High	34,600	0.44	99.6	49.0
Medium	49,885	0.64	99.0	70.3
Low	51,976	0.66	98.6	72.9
Very Low	57,790	0.74	94.9	78.1

* The total number of CDR records is 78,349.

8. DISCREPANCIES IN REPRESENTATION OF SUBGROUPS

At first one may think that link accuracy is far more important than match-link rate; and for certain applications this would be true. A higher link accuracy at the expense of the match-link rate is obtained when a high cut-off is set for linking. Record-pairs with comparison weights above the high cut-off will tend to have good agreement on all or most of the linking variables. If such record-pairs are randomly distributed throughout the population, then the decision to require high accuracy may well be valid.

However, it is more likely that some subgroups within the population will have a smaller chance of meeting these stringent requirements. It is worth comparing the different standards and levels of linked datasets to see how different criteria used in linking affect the proportions of various subpopulations present. For a given variable, each linked dataset was divided into categories and the proportion of the linked records in each category in the CDR was calculated. This procedure was conducted for a range of linking and analytical variables.

Two linking variables that show large changes among the linked datasets are age (figures A.1, A.2, A.3 in the Appendix) and Indigenous status (figure A.4). Younger people aged 0–19 and Indigenous people are under-represented in the Bronze Standard. Under-representation is more pronounced for the Bronze Standard than for the Silver Standard and, in both standards, the under-representation trend increases as the cut-off increases.

Young people who were born in Australia have very few extra variables present to enable them to be distinguished from each other. They may not have finished schooling, they may not yet have qualifications and most will be single. If, in addition, they report their age but not their date of birth there is even less to distinguish them. Examples of the effect of not having extra characteristics for linking is shown in tables A.7 and A.8. Only 7% of Bronze High links have years of schooling either missing or designated not applicable compared with 26.3% in the Gold Standard. In addition, 55% of Bronze High links have post-school qualifications either missing or not applicable compared with 68.5% of Gold links.

Indigenous persons may also have higher levels of missing linking variable values but an issue that arose in this quality study, and is expected to improve when linking the SLCD, is that of mesh blocks. During the Census Dress Rehearsal, coding to mesh blocks in some regional areas, including discrete Indigenous communities, was not very accurate. Thus Indigenous people from remote communities are more likely to have a missing or imputed mesh block on the Census Dress Rehearsal. In the Bronze Standard linkages, the importance of this geographic variable is shown by the fact that very few Indigenous persons from remote communities were linked in this standard.

Extra resources were applied in the 2006 Census resulting in some improvement and there will be further improvements in 2011, through the use of additional administrative data.

Other linking variables that show discrepancies among the different linked datasets are Marital status, Country of birth, Highest level of schooling completed, Level of education and Year of arrival (tables A.5, A.6, A.7, A.8, A.9). People born overseas (particularly those who arrived more than 10 years ago), more highly educated people and married people are over-represented in Bronze Standard linkages. This under- and over-representation is much weaker or non-existent in Silver Standard. Sex and State of usual residence are represented consistently for all standards.

Some analytical variables that show differences among linked datasets are Occupation, Industry of employment and Income (tables A.10, A.11 and A.12 respectively). People employed in Agriculture, Forestry and Fishing are slightly under-represented in Silver and much more so in Bronze, whereas those employed in Financial and Insurance Services, Health Care and Social Assistance, Manufacturing, Professional, Scientific and Technical Services, and Public Administration and Safety, are all slightly over-represented in the Bronze Standard with a less-marked difference in the Silver Standard. A contributing factor to the under-representation of those employed in Agriculture, Forestry and Fishing is the higher rate of missing and imputed mesh block codes in rural areas, as mentioned earlier.

There are some variations in representation of the occupation groups. It is particularly interesting to compare Silver and Bronze Very Low with Gold. For instance, Labourers are under-represented in both but much more so in Bronze, while Managers are under-represented in Bronze but not in Silver. Professionals and Clerical and Administrative Workers are slightly over-represented in Silver and more so in Bronze.

Of those people who responded to the income question in 2005, people with higher incomes are over-represented and people with lower incomes under-represented in both the Bronze and Silver Very Low compared with Gold, with the effect more marked for Bronze than Silver. This is commensurate with the findings for Level of education and Occupation.

Generally the ability of a linkage to represent the CDR population is ordered from best to worst as follows: Gold Standard, Silver Standard (Very Low, Low, Medium, High), Bronze Standard (Very Low, Low, Medium, High) although, for some variables, Bronze Very Low Standard may be better than Silver High. Subsequent investigations have focused on Gold, Silver Very Low and Bronze Very Low Standards.

Of particular interest is the fact that the Silver Standard better represents the subpopulations of young people and of Indigenous people than the Bronze Standard.

9. PERFORMING ANALYSES WITH LINKED DATA

While it is important to consider the issues raised in Sections 7 and 8, it is also of interest to see what effect discrepancies in representation and the quality measures (link accuracy and match-link rate) are likely to have on some typical analyses. Several sources of potential research questions that might be asked of longitudinally linked Census data, were considered.

- At the Census Analysis Conference (July 2006), Census data users posed various issues that could be explored with the data if they had information about each person at the previous Census.
- At the Household Income and Labour Dynamics in Australia (HILDA) Survey Users Conference (July 2007), a number of different issues using longitudinal HILDA data were considered. Many of these could be examined for smaller population subgroups with the much larger linked Census dataset.
- The Office for National Statistics in the United Kingdom has been linking Census data since the 1971 Census and some of the research projects conducted on that dataset could be applied in an Australian context.
- Finally, an expert group was assembled in the ABS for a roundtable discussion in February 2008 to discuss the sorts of questions being considered for analysis and the way in which the ABS was conducting the analyses.

While one focus of the analysis of linked Census data, collected at five-yearly intervals, might be to examine the effects of government policy changes, such an analysis is unlikely to be fruitful with linked CDR and Census data, collected only one year apart. Consequently our focus has been on variables that are likely to change over one year without government intervention.

It should be noted that, as the CDR is not a randomly collected sample, any findings from the analyses reported in the rest of this section cannot be extrapolated to the Australian population. They are presented to show the effects of data linkage standards and levels on analytical results.

9.1 Bivariate analyses

We have considered two analyses in tabular form. First, of interest to people planning resources for teachers, was to consider what proportions left this occupation to pursue alternative occupations rather than leaving to retire. There were 288,707 (1.5%) school teachers in the 2006 Census file and 1,157 (1.6%) Gold Standard linked records where occupation was listed as School Teacher in either 2005 or 2006 or both. The numbers are too low for an in-depth modelling exercise using the linked data but can be presented in a tabular format. An example of comparative analysis for Gold

Standard, Bronze Very Low and Silver Very Low is shown in table A.13 of the Appendix. From this rudimentary analysis there is no obvious difference in conclusions among the three standards. The slight over-representation of teachers over 55 years and under-representation of teachers under 25 is consistent with the findings in Section 8.

The second analysis, presented in table A.14 of the Appendix, relates to Indigenous employment. It examines changes in employment status between 2005 and 2006 for Indigenous people. Using the Gold Standard data, 79.8% of Indigenous persons employed in 2005 were still employed in 2006 while the figures for Silver and Bronze Standards were 86.9% and 88.7%, respectively. The sample sizes are fairly small with employment status recorded for only 987 linked indigenous pairs in Gold Standard, 756 in Silver and 567 in Bronze. However, the confidence interval for the Gold Standard is (74.9%, 82.1%) and the Silver and Bronze estimates lie outside this interval. This result is consistent with the general pattern of over-representation of employed people in Silver and Bronze Standards and the under-representation of Indigenous people in those standards.

9.2 Fitting models to linked data

Several analyses were conducted to address the issue of whether different conclusions would be reached if the Bronze Very Low, Silver Very Low or Gold Standard linked datasets were used in model fitting. Linked data were used to fit three logistic regression models:

1. the odds that a person aged 15 or more is employed in 2006 as a function of 2005 explanatory variables;
2. the odds that a person moves between 2005 and 2006 as a function of 2005 explanatory variables; and
3. the odds that a person aged 15 or more is a student in 2006 as a function of 2005 explanatory variables.

A multiple linear regression model was used to model a change in hours worked between 2005 and 2006 for people aged 15–54 who were employed in at least one of the two years.

A measure of deviance was used to compare coefficients for each model across datasets. It is given by the following formula (Chipperfield, 2009):

$$D = \frac{1}{K} \sum_{k=1}^{K} \frac{\left| G_k - S_k \right|}{se\left\{ G_k \right\}}$$

where

S_k is the k-th model parameter for standard S,

G_k is the k-th model parameter using the Gold Standard,

$se\{G_k\}$ is the standard error of the k-th model parameter using the Gold Standard,

and $k = 1, \ldots, K$, where K is the number of coefficients in the model.

Measures of deviance for two models are presented in table 9.1. A smaller value for D indicates the fitted model more closely resembles the model fitted to the Gold Standard linked data. The trend is for models to more closely resemble the Gold Standard model as the cut-off is lowered. However, as shown in table 9.1, change in hours worked breaks the trend for the Very Low cut-off.

9.1 Deviance (*D*) measuring the average standardised difference between model coefficients for Gold Standard and other standards of linked data

	Bronze Standard				*Silver Standard*		
Model	*High*	*Medium*	*Low*	*Very Low*	*High*	*Medium*	*Low*
Changes in hours worked from 2005 to 2006	0.92	0.82	0.71	0.83	0.52	0.40	0.42
Odds that a person moves between 2005 and 2006	1.23	0.80	0.72	0.58	0.61	0.46	0.40

These findings led us to refine our modelling methods so that datasets did not inadvertently target under-represented population groups and to focus our attention on Very Low cut-offs.

Table 9.2 shows the explanatory variables used in fitting each of the three logistic regression models. On the whole there was good agreement among models fitted to the three linked datasets, with each dataset usually having the same explanatory variables included as significant. Occasionally there were differences, such as Indigenous status being significant in explaining the odds of employment for the model fitted using Gold Standard linked data, but was not for the corresponding models fitted using Bronze Very Low and Silver Very Low linked data. This could be caused by the relatively small number of Indigenous persons with complete sets of explanatory variables in the Bronze Very Low (362) and Silver Very Low (489) datasets compared with the Gold Standard (630) linked data. It could be also be a result of the differential linkage rate for Indigenous persons; those in remote areas are less likely to be linked than those in regional and urban areas. To make comparison among datasets meaningful, all variables were forced into the models.

9.2 Explanatory variables from 2005 used in fitting logistic regression models to linked data

Model 1 *Odds that a person is employed in 2006*	*Model 2* *Odds that a person moves between 2005 and 2006*	*Model 3* *Odds that a person is a student in 2006*
	Personal characteristics	
Female	Female	Female
Indigenous		Indigenous
	Not married	Not married
Required disability assistance		Required disability assistance
Moved usual residence in last year	Moved usual residence in last year	
Age	Age	Age
Aged 55–69 years	Aged 15–24 years	Aged 25–39 years
Aged 70 years or over	Aged 40–54 years	Aged 40–54 years
	Aged 55–69 years	Aged 55–69 years
	Aged 70 years or over	Aged 70 years or over
Tenure status	Tenure status	
Paying off house	Paying off house	
Renting	Renting	
	Work characteristics	
Employed		
Unemployed		
Hours worked	Hours worked	Hours worked
Provided unpaid disability care		Provided unpaid disability care
Volunteer		Volunteer
Income per week	Income per week	Income per week
$1–149	$1–149	$1600–1999
$150–249	$150–249	
$250–399		
$400–599		
$600–799		
$800–999		
$1000–1299		
$1300–1599		
$1600–1999		
Occupation	Occupation	Occupation
Labourer	Machinery operator and driver	Labourer
Sales or Retail	Salesperson	Machinery operator and driver
	Manager	Manager
	Community and personal service worker	
	Education characteristics	
	High school student	High school student
TAFE student		TAFE student
University student	University student	University student
Had a degree or equivalent	Had a degree or equivalent	Had a degree or equivalent

While similarity of fitted models is important, it is also worthwhile considering confidence intervals for test cases predicted by the models. Five test cases were selected and are shown in table 9.3 with the rationale for their selection.

9.3 Test cases used for obtaining predictions from each of the three logistic regression models

Test case	*Description (in terms of 2005 CDR variables)*	*Rationale*
1. Female child carer	• provided unpaid child care • worked part-time for 15 hours a week as a salesperson • was female • was aged 25–39 • earned $400–$599 a week • lived in a dwelling which was being purchased • was in the base category of all other variables	Policy driven rationale: • on 1 July 2006, the baby bonus increased • in the Budget announced in May 2006, it was announced that, from 1 July 2006, eligibility for the maximum rate of Family Tax Benefit Part A would be extended.
2. Male Indigenous person seeking employment	• Indigenous • earned $250–$399 a week • actively seeking work (full or part-time) • was in the base category of all other variables	Univariate investigations concluded that young and Indigenous people were under-represented in the Silver and Bronze Standards.
3. 19 year-old student	• worked part-time for 12 hours a week as a salesperson • earned $250–$399 a week • was aged 15–24 • lived in a dwelling which was being rented • was a university student • was in the base category of all other variables	Univariate investigations concluded that young people were under-represented in the Silver and Bronze Standards.
4. Young male Indigenous not in the labour force	• was aged 15–24 • was not in the labour force • was in the base category of all other variables	Univariate investigations concluded that young and Indigenous people were under-represented in the Silver and Bronze Standards.
5. Typical case	• male • non-Indigenous • married • 40 hours worked each week • aged 25–39 • did not move in the previous year • Professional occupation • did not possess a degree • owned own house • earned $600–799 a week	Designed to be a well-represented subpopulation for testing purposes.

The confidence intervals for the probability of being employed in 2006 obtained from the first logistic regression model are shown in figure A.15 of the Appendix for illustration purposes. In general, the predicted probability of an event for a test case did not vary greatly whether the models were fitted using Gold, Bronze Very Low or Silver Very Low data.

The discriminatory power of the models was consistent across the Gold, Bronze Very Low and Silver Very Low Standards. In particular, all models have good or excellent ability to discriminate among those subpopulations which are more or less likely to move address, be employed or be a student in 2006. It should be noted that goodness of fit diagnostics for Gold, Bronze Very Low and Silver Very Low models suggest that none of the logistic models provide a reasonable fit to some subpopulations, causing an unusual pattern in the residuals. This is not unusual when modelling social variables, such as those used here.

The above arguments lead to the conclusion that one would make similar conclusions whether the model was fitted to Gold, Bronze Very Low or Silver Very Low linked data. We note, however, that these conclusions may be particular to these models.

10. MODIFICATIONS TO FITTED MODELS

10.1 Adjusting models for inexactly linked data

As has been shown in Section 7, some linked record-pairs do not belong to the same person. If analyses are performed on the linked dataset as if all links were correct, the resulting regression coefficients will in general be biased towards zero. Chipperfield (2009) describes the implementation of a method to adjust regression coefficients when linkage is inexact. He was able to demonstrate that his implementation made corrections for errors in the linkage. However, this source of error was overshadowed by the bias in regression estimates caused by over- or under-representation of some subgroups in the linked data because of missed links. The subpopulations affected are discussed in detail in Section 8.

10.2 Weighted analyses

One way to overcome issues of under-representation is to use weighted analyses. The difficulty for this quality study lay in finding an appropriate benchmark. The Census Dress Rehearsal is not a random sample of the Census population and so Census counts are not appropriate to use as benchmarks. They would, however, be appropriate benchmarks for the SLCD when it is linked with the next Census as the SLCD is a random sample of the Census. For lack of any other alternative we have used the Census Dress Rehearsal counts as benchmarks.

Weighting variables are shown in table 10.1.

10.1 Benchmark variables from CDR used to weight linked data

Variables	*Categories*
Indigenous_status_missing	{present, missing}
Marital_status	{married, never_married, divorced, widowed, separated, NA}
Indigenous × Age	{non-Indigenous or missing, Indigenous} × {0–14, 15–24, 25–34, >35}
Year_of_arrival	{2002–05, 1998–2001, <1998, missing, NA}
Born_os × Sex × Age	{Australian-born or missing, overseas born} × {M,F} × {0–14, 15–24, 25–34, 35–64, 65–79, >80}

Table 10.2 compares the percentage of each of a number of characteristics in Bronze Standard linked records for weighted and unweighted data with the percentage in Gold Standard linked data. From this we can see that under- and over-representation

are improved for six of the ten characteristics displayed. One characteristic, Employed in Agriculture, is changed little by the weighting while the remaining three were made worse; for instance, Never married.

10.2 Percentages of various characteristics in weighted and unweighted data compared with the Gold Standard and the full Census Dress Rehearsal

	Bronze Standard – Very Low			
	Unweighted	*Weighted*	*Gold Standard*	*CDR*
Never married*	30.04	32.32	30.79	32.08
Previously married*	16.15	17.24	17.01	17.46
Currently married*	53.80	50.43	52.14	50.29
Indigenous*	1.62	3.26	2.42	3.20
Speaks English at home	73.26	74.58	74.15	72.09
Year arrival before 1997*	21.72	18.75	18.99	18.67
Born overseas*	31.94	29.03	28.82	29.64
Household income < $400	23.29	24.02	24.58	26.05
Industry Agriculture	2.14	2.13	3.57	3.51
Year 12	40.82	38.88	37.74	37.96

* Used in some form in the weighting.

10.3 Forcing linking variables into the model

Linking variables may be influential in a model even if they are not statistically significant, as they may aid in adjusting for under- or over-representation of groups with particular characteristics. If linking variables are included as covariates in a model they may help overcome bias caused by missed links. Two approaches were taken to force linking variables into the three models discussed in Section 9.

In the first approach, the linking variables included in forced models were selected according to their role in the linking process, and some possibility of over- or under-representation on Silver or Bronze. As younger individuals were more difficult to link, an age category capturing status as a 15–24 year old was forced into the model (under 15s were excluded from the models). Religion was used as a linking variable, and there were some differences in representation of various groups. Accordingly four binary religion variables were formed, *viz.* Catholic, Anglican, Other Christian, and Non-Christian, with No religion acting as the base case.

The complete set of variables forced into the models is shown in the left column of table 10.3.

10.3 Linking variables forced into models

Linking variables from 2005	*Binary variables indicating missing linking variable values from 2005**
Language other than English spoken at home	Language spoken at home missing
Born overseas	Country of birth missing
Highest year of schooling year 12	Highest level of schooling missing
Highest year of schooling years 10 or 11	Religion missing
Catholic	Mesh block missing
Anglican	Field of highest qualification missing
Other Christian	Level of highest qualification missing
Non-Christian religion	
Rural region	
Remote region	
Level of highest qualification post graduate degree	
Level of highest qualification bachelor degree	
Level of highest qualification diploma	
Sex	
Indigenous status	
Binary marital status (married/not married)	

* Age and Sex were never missing

The aim was to compare fitted models using the deviance measure in Section 9.2. However, it is difficult to make direct comparisons between models fitted to two different linkage standards if different variables are statistically significant in each model. One method was to calculate the deviance from variables that were statistically significant in both the Gold Standard and the other standard.

The deviances were found to increase slightly rather than to decrease, with the forced inclusion of linking variables. This was in part due to a reduction in the number of complete records when additional variables were included.

For this reason, the second approach was to force missing flag variables representing the missing status of values of linking variables into the model. This approach would not lead to the dropping of any extra records because of missing linking variable values. Furthermore, missed links are more likely to be caused by missing linking variables rather than the actual value of a linking variable that is present.

Missing flag variables that were included are shown in the right column of table 10.3. In each case a value of 1 indicates the variable was missing for a given record, while a value of 0 indicates a value for the variable was present on the record. Age and Sex were never missing from records as they were imputed in Census processing if they were not completed on the form. There was no variable available to indicate missing status for hash value.

To make comparisons easier, missing flag variables were also forced into the model fitted to the Gold Standard. Forcing flag variables resulted in different explanatory variables being included in final models. To calculate the deviance for a Bronze Standard model, say, only variables that were included in both the Bronze and Gold models could be compared and so variables that were not significant in one or the other were not included in the calculation.

Table 10.4 shows that the deviance measure decreased for each of the three models and for both Bronze and Silver Standards when missing linking variable flags were included. In all cases the decrease was less than 10% but consistent, suggesting that this method may be useful in correcting some of the bias caused by missed links.

10.4 Deviance measures comparing the effect of forcing missing linking variable flags into model

	Missing flag variables omitted		*Missing flag variables forced*	
	Silver Standard	*Bronze Standard*	*Silver Standard*	*Bronze Standard*
Odds that a person is employed in 2006 (model 1)	0.507	0.665	0.461	0.614
Odds that a person moves between 2005 an 2006 (model 2)	0.491	0.828	0.468	0.784
Odds that a person is a student in 2006 (model 3)	0.508	0.751	0.494	0.729

11. LIKELY DEGRADATION OF DATA QUALITY OVER FIVE YEARS

The five-year gap between successive Censuses will lead to more changes in variable values for an individual than have been seen between the 2006 Census and CDR with their one-year gap. The best linking variables, such as Date of birth, Country of birth and Year of arrival in Australia, do not change over time but people's reporting of them may change a little. Some variables that have been useful in linking, such as Highest level of schooling, Level of highest qualification and Field of study of highest qualification, are quite likely to change, particularly for younger people.

The probabilistic linkage method allows for changes in an individual's values between two datasets by altering the input parameters, i.e. the m- and u-probabilities. Additional change over a five-year period could be included by decreasing the input m-probabilities. See Conn and Bishop (2006) for a description of the probabilities and Solon and Bishop (2009) for the values of the probabilities that have been used in the linkage of CDR to Census.

One important linking variable is mesh block. For this simulation, we have used mesh blocks derived from CDR usual address and usual address one-year-ago, as reported on the Census form. When linking 2006 and 2011 Censuses, mesh blocks that have been derived from the usual address as reported in the 2006 Census, and the usual address five-years-ago as reported in the 2011 Census will be used. Problems may arise if, in 2011, people forget where they usually lived five years beforehand.

It is possible to assess likely degradation in mesh block quality over five years and obtain an indication of the resulting degradation of match-link rates, by using five-years-ago address reported in the 2006 Census. The Census form asked respondents to provide their address five years ago if it was different from one of the addresses they have provided already, namely address one year ago, usual address, or front of form address.

By using published Census counts we were able to estimate the number of SLCD persons whose five-year-ago address would be different from the other addresses. Call this n. A random sample of 50,000 people was selected from the first wave of the SLCD and any distinct five-year-ago addresses were mesh block coded. The proportion, p, of distinct five-year-ago addresses that could not be coded because of inadequate address information was calculated. For example, if the address given is "Sydney", this would not be coded.

The two quantities n and p were multiplied to determine the additional number of matches that might not be linked due to the deterioration of mesh block coding of five-year-ago addresses (np). This is a worst case scenario, implying that other variables would not be good enough to the make the links. This number (np) was

then used to adjust the match-link rate achieved in the Bronze Standard Very Low linkage, which is the most likely method to be used for the planned linking of the 2006 wave of the SLCD with the 2011 Census.

The match-link rate for the Bronze Very Low linkage was 79%. The proportion of distinct five-year-ago address that could not be mesh block coded was 12%. Applying this to expected proportion of people that would move between one and five years allowed for an adjustment to match-link rate from 79% to 77%. These results are shown in table 11.1.

11.1 Possible effect of mesh block quality degradation over five years (worst case scenario)

Scenario	*Match-link rate*	*Number of linked record-pairs*
Complete SLCD linkage	100%	979,794
Current match-link rate (Bronze Very Low)	79%	774,037
Missing mesh block for five-years-ago address prevents linkage		–19,596
Revised match-link rate	77%	754,441

12. CONCLUSIONS

Linking the 2006 Census with the Census Dress Rehearsal conducted one year earlier has been a useful simulation to investigate the methodological, process and quality issues in forming the SLCD.

As discussed in Section 6, the Gold Standard linked dataset is of high quality and acts as a useful benchmark for the other standards of linkage. The Bronze Standard linkage represents the method that is likely to be used for the planned linking of the 2006 wave of the SLCD with the 2011 Census. A Very Low cut-off will be the most likely choice, given the various findings reported above and, particularly, that missed links cause more bias than incorrect links in analyses that use linked data. The Silver Standard linkage has been investigated as a possibility for later waves but this would only proceed subject to favourable response from both further public consultation and a privacy impact assessment.

One should be mindful of the fact that this is a simulation and not the real exercise. While the findings reported in this paper can largely be extrapolated to the formation of the SLCD, there will be some differences. The most notable will be in the differences in data quality between the Census Dress Rehearsal and the Census.

Investigations, reported in Section 8, have shown some population groups are under-represented when linking without name and address. This is mainly due to the fact that other information must be used to establish the links; subpopulations which do not have reliable other information are less likely to be linked. Under-represented subpopulations are likely to have limited address information, and hence missing or imputed mesh block codes, or lack additional distinguishing characteristics, such as country of birth outside Australia or a post-school qualification. It is useful to compare the quality of the Census Dress Rehearsal with the Census to gauge whether these problems encountered in this simulation will persist when linking the SLCD.

There are proportionately more missing values for some linkage variables in the Census Dress Rehearsal than in the Census. Those variables with higher rates of missingness include Level of post-school qualification, Field of study of post-school qualification, Highest year of school completed and Indigenous status.

However, most important is the higher proportion of missing mesh blocks in the Census Dress Rehearsal. Mesh blocks for the 2006 Census were experimental, as indicated in an information paper (ABS, 2008a) and were even more experimental for the Census Dress Rehearsal. More resources were also expended on the Census data for manual checking of incomplete addresses resulting in more complete mesh block coding. The more complete and robust introduction of rural addressing standards in non-metropolitan areas, principally for emergency management purposes, will improve the degree of mesh block coding in 2011.

Assuming that the level of missingness of linkage variables in the 2011 Census is similar to that in the 2006 Census, the variables for linking the two Censuses will be of higher quality than those used in the simulation reported here. Not only will this result in a higher match-link rate than was obtained for the Bronze Standard, but also better mesh block coding may mean some of the under-represented groups in the simulation will be less likely to be under-represented in the SLCD. Those particularly likely to experience improved representation are Indigenous people in remote communities and those employed in agriculture.

There may still be some under-representation of subpopulations, particularly young people. Our analyses of some typical questions, reported in Section 9, have shown that one would not reach very different conclusions for the different standards of linkage. However, there are some important caveats. For instance, if the focus of analysis is on children, or following children into adulthood, there may be some problems, as young persons are under-represented in the Bronze Standard linkage. Weighting to the full Census may help to overcome this issue. Forcing linking variables into a fitted model may also help.

To overcome bias in regression coefficients when fitting models to the linked data, the method mentioned in Section 10.1 will be useful and work is continuing in this sphere.

The degradation of mesh block quality over a five-year period has been shown to have some effect on match-link rate, in Section 11. While the extent of change in other linking variables over five years is not completely known, it should be possible to extrapolate from single year changes and make necessary adjustments to the m- and u-probabilities provided as input in the linking.

This study has shown that the inclusion of hash values in the set of linkage variables, as used in the Silver Standard in the simulation, would overcome most of the problems with under-representation of subpopulations. Add to this the improvements in match-link rate and link accuracy and the method for adjusting fitted model coefficients, then the linkage represented by the Silver Standard would perform very well.

The main concerns of forming the SLCD without using name and address are under-representation of certain subpopulations and incorrect linkages. To put these concerns in perspective, consider the properties of a high quality longitudinal household survey, namely the Household Income and Labour Dynamics in Australia (HILDA) survey. This is a household-based panel survey which began in 2001 run by the Melbourne Institute of Applied Economic and Social Research on behalf of the Department of Families, Housing, Community Services and Indigenous Affairs.

Some properties of the HILDA survey prevent direct comparison with the SLCD. For instance, HILDA samples households and when these break up, the survey follows all members to their new households. On the other hand the SLCD samples persons. Each person's information can include such aspects as total household income or family type but details of other people in the household are not included.

In the first wave of the HILDA survey, the initial response rate was 66% of sampled households; in these there were 15,127 eligible members of whom 13,969 were interviewed (Watson and Wooden, 2006). Assuming similar composition between responding and non-responding households, there would have been 23,025 persons in the original sample, giving an overall response rate of 61%. In a panel survey, there are no incorrect links from one wave to another as the respondent's identity is known completely.

The official Census night figure was 19,855,288 from which a 5% random sample was to be selected. This figure includes adjustments for net undercount and imputed persons, explained in other ABS publications (2007c and 2008b, respectively). Only 19,050,146 Census records were available from which to select the sample. This is 95.9% of the target population and can be considered as similar to an initial response rate when comparing with surveys. The available records were slightly over-sampled to compensate for the lesser numbers, yielding the first-wave of the SLCD consisting of 979,794 persons. Using the figures obtained from this quality study as an approximation to match-link rate and link accuracy for the first two waves of the SLCD, HILDA and the SLCD are compared in table 12.1.

12.1 Comparison of the HILDA panel survey after five years and possible SLCD after two Censuses five years apart. The Silver Standard is included for illustration purposes only.

		Statistical Longitudinal Census Dataset	
	HILDA	*Bronze Standard Very Low*	*Silver Standard Very Low*
Initial response rate	66% of households × 92% of persons/household = 61% of persons	95.9% of persons	95.9% of persons
Response after 5 years	74%	78.1%	91.3%
Number of waves	5	2	2
Incorrect links	0	5.1% *	3.7% *
Sample size	13,969	979,794	979,794

* 100% – Link accuracy

In summary, the biggest problem with linking data to form the SLCD will be missed links but this is unlikely to be on as large a scale as non-response in household panel surveys.

REFERENCES

Australian Bureau of Statistics (2005a) *Census Data Enhancement – Statement of Intention*, (last viewed on 5 August 2009) <http://www.abs.gov.au/websitedbs/D3110124.NSF/f5c7b8fb229cf017ca256973001fecec/5812a287d6a2e78fca2571ee001a7a49!OpenDocument>

—— (2005b) *Enhancing the Population Census: Developing a Longitudinal View*, Discussion Paper, cat. no. 2060.0, ABS, Canberra.

—— (2006a) *Census Data Enhancement Project: An Update*, Information Paper, cat. no. 2062.0, ABS, Canberra.

—— (2006b) *How Australia Conducts a Census*, Information Paper, cat. no. 2903.0, ABS, Canberra.

—— (2007a) *Review of the Australian Standard Geographical Classification, 2007*, Information Paper, cat. no. 1216.0.55.001, ABS, Canberra.

—— (2007b) *Australian Demographic Statistics, March 2007*, cat. no. 3101.0, ABS, Canberra.

—— (2007c) *Census of Population and Housing – Details of Undercount, August 2006*, cat. no. 2940.0, ABS, Canberra.

—— (2008a) *Outcomes from the Review of the Australian Standard Geographical Classification,* Information Paper, cat. no. 1216.0.55.002, ABS, Canberra.

—— (2008b) *Census Dictionary, 2006*, cat. no. 2901.0, ABS, Canberra.

—— (2009) *Concepts Sources and Methods: Population Estimates, 2006*, cat. no. 2901.0, ABS, Canberra

Bishop, G. and Khoo, J. (2007) "Methodology of Evaluating the Quality of Probabilistic Linking, *Methodology Research Papers*, cat. no. 1351.0.55.018, Australian Bureau of Statistics, Canberra.

Chipperfield, J.O. (2009) "Generalised Linear Models with Probabilistically Linked Data", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.098, Australian Bureau of Statistics, Canberra.

Conn, L. and Bishop, G. (2006) "Exploring Methods for Creating a Longitudinal Census Dataset", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Solon, R. and Bishop, G. (2009) "A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.

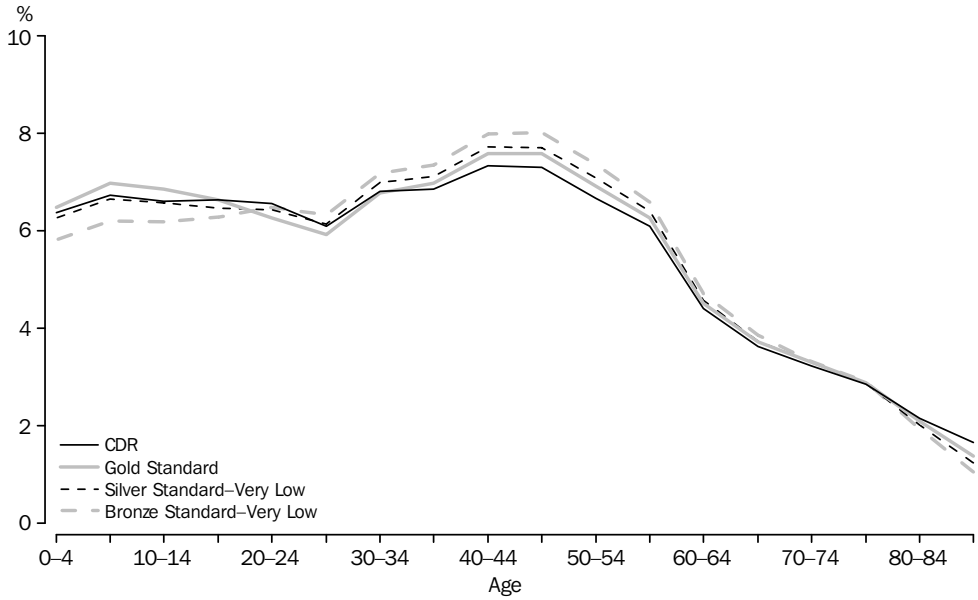
Watson, N. and Wooden, M. (2006) *Modelling Longitudinal Survey Response: The Experience of the HILDA Survey*, HILDA Project Discussion Paper Series, no. 2/06.

ACKNOWLEDGEMENTS

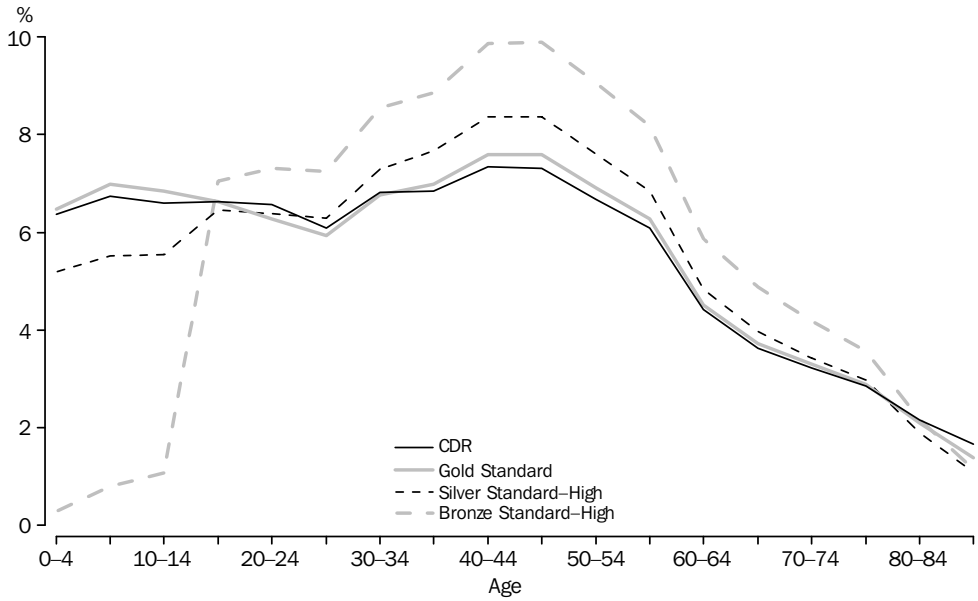
The author wishes to acknowledge the analytical work of James Chipperfield, Tenniel Guiver, Richard Solon, Paul Campbell, Tim Ayre and Shaun McNaughton, without which this paper would not have been possible. The author would like to thank Peter Rossiter, Alan Wong and Marcus Blake for their valuable comments.

APPENDIX

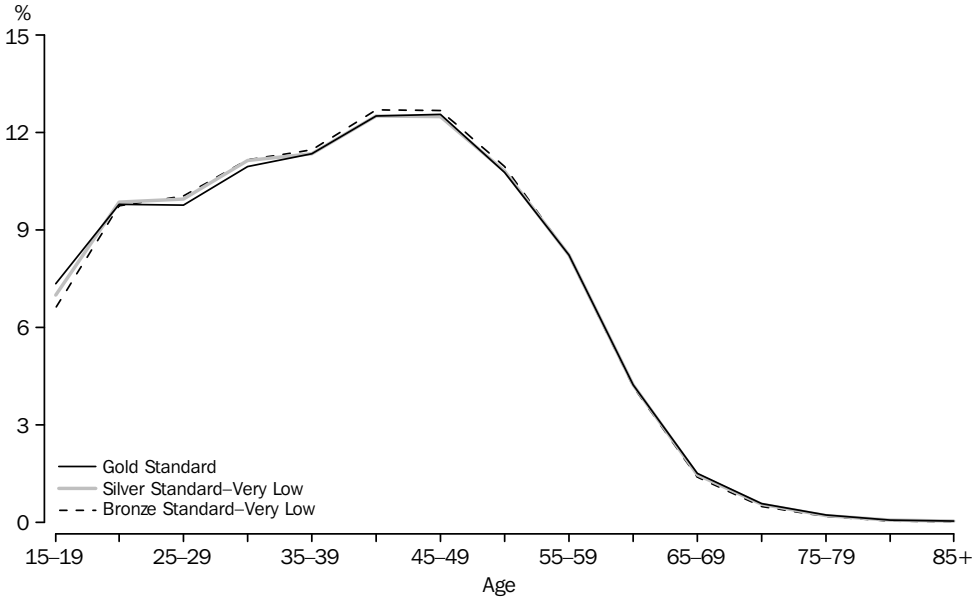
A.1 Comparison of age distributions for the CDR, Gold Standard, and Silver and Bronze Very Low Standards



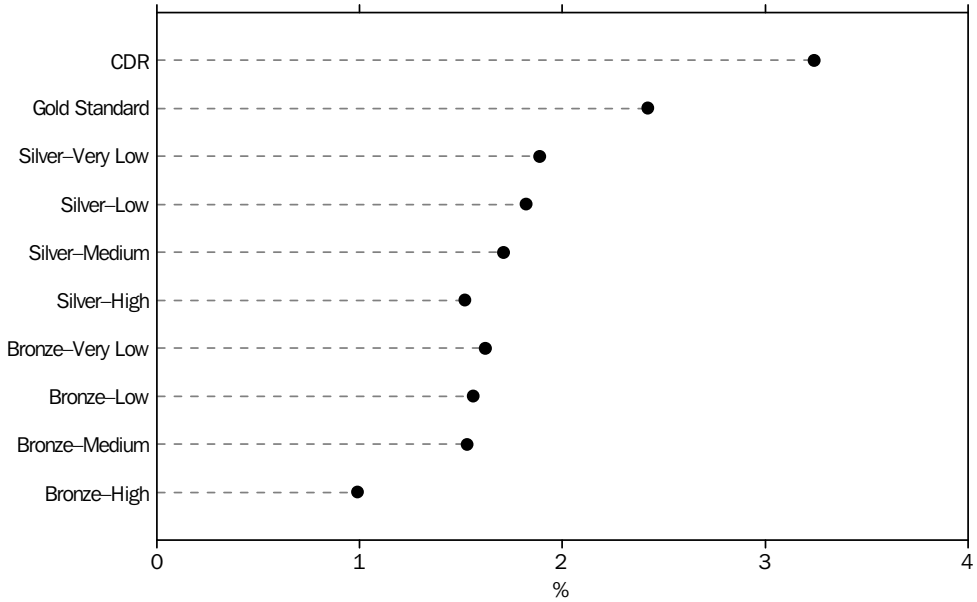
A.2 Comparison of age distributions for the CDR, Gold Standard, and Silver and Bronze High Standards



A.3 Age distribution for people employed in at least one of 2005 and 2006



A.4 Percentage of linked people who identify as Aboriginal and/or Torres Strait Islanders for all linkage standards



A.5 Distribution (%) of Marital status* – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages. (Records with this item missing have been excluded)

	Marital status		
	Married	*Never married*	*Separated / divorced / widowed*
CDR	50.4	32.1	17.5
Gold Standard	52.2	30.8	17.0
Silver Standard			
Very Low	52.8	30.7	16.6
Low	53.0	30.2	16.7
Medium	53.4	30.1	16.6
High	54.5	29.5	16.0
Bronze Standard			
Very Low	53.8	30.0	16.1
Low	54.3	29.3	16.4
Medium	54.3	29.3	16.3
High	55.8	28.4	15.7

* People aged 15 years and over only.

A.6 Distribution (%) of Country of birth – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages

	Country of birth					
	Australia	*Asia*	*Europe*	*UK & Ireland*	*New Zealand*	*Other*
CDR	70.2	12.4	5.5	4.7	1.6	5.7
Gold Standard	71.0	12.1	5.3	4.5	1.5	5.6
Silver Standard						
Very Low	70.2	12.4	5.5	4.8	1.5	5.7
Low	70.1	12.3	5.5	4.8	1.5	5.7
Medium	69.8	12.5	5.6	4.8	1.5	5.8
High	67.7	13.5	6.0	5.2	1.6	6.2
Bronze Standard						
Very Low	67.9	13.5	5.8	5.2	1.6	6.0
Low	67.4	13.6	6.0	5.3	1.6	6.2
Medium	68.2	13.2	5.8	5.2	1.5	6.0
High	60.3	16.4	7.4	6.5	1.8	7.5

A.7 Distribution (%) of Highest level of schooling completed – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages

	Highest year of school						
	No school	*Year 8 or below*	*Year 9*	*Year 10*	*Year 11*	*Year 12*	*NA/ Missing*
CDR	0.8	6.2	5.8	17.9	7.2	35.5	26.7
Gold Standard	0.7	6.0	5.8	18.4	7.3	35.6	26.3
Silver Standard							
Very Low	0.7	5.8	5.7	18.4	7.4	37.0	25.0
Low	0.7	5.9	5.8	18.5	7.3	36.8	25.0
Medium	0.6	5.9	5.9	18.8	7.4	37.4	24.0
High	0.6	6.0	5.9	19.3	7.6	39.8	20.9
Bronze Standard							
Very Low	0.7	5.8	5.6	18.5	7.5	38.8	23.2
Low	0.7	5.9	5.7	18.4	7.4	38.3	23.7
Medium	0.6	5.9	5.8	18.5	7.5	38.1	23.7
High	0.7	7.0	6.9	22.2	8.8	47.5	7.0

A.8 Distribution (%) of Level of education – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages

	Level of education				
	Postgraduate	*Bachelor*	*Technical*	*NA*	*Missing*
CDR	3.5	10.2	17.4	59.4	9.6
Gold Standard	3.5	10.2	17.8	59.8	8.7
Silver Standard					
Very Low	3.7	11.0	19.0	58.0	8.4
Low	3.7	10.7	18.5	58.9	8.2
Medium	3.8	11.0	18.9	58.2	8.2
High	4.2	12.0	20.6	55.5	7.7
Bronze Standard					
Very Low	4.0	12.0	20.3	55.9	7.9
Low	3.9	11.6	19.4	57.1	8.0
Medium	3.9	11.5	19.3	57.5	7.8
High	5.1	14.7	25.2	46.5	8.5

A.9 Distribution (%) of Year of arrival – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages

	Year of arrival			
	up to 1996	*1997–2001*	*2002–2006*	*NA/ Missing*
CDR	18.8	4.4	3.6	73.2
Gold Standard	19.0	4.3	3.1	73.6
Silver Standard				
Very Low	19.9	4.4	3.3	72.4
Low	20.2	4.4	3.2	72.2
Medium	20.5	4.5	3.2	71.9
High	22.2	4.9	3.4	69.5
Bronze Standard				
Very Low	21.7	4.8	3.5	70.0
Low	22.4	4.8	3.4	69.4
Medium	22.0	4.7	3.2	70.1
High	28.1	5.8	4.0	62.0

A.10 Distribution (%) of Occupation (1-digit groups) – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages

	Occupation group			
	Clerical & Administrative Workers	*Community & Personal Service Workers*	*Labourers*	*Machinery Operators & Drivers*
CDR	15.3	8.3	11.9	7.5
Gold Standard	15.5	8.1	11.8	7.4
Silver Standard				
Very Low	15.5	8.2	11.4	7.3
Low	15.8	8.1	11.5	7.4
Medium	15.8	8.1	11.5	7.4
High	15.8	8.0	11.1	7.3
Bronze Standard				
Very Low	15.8	8.0	11.0	7.4
Low	16.2	8.0	11.3	7.6
Medium	16.3	8.0	11.4	7.6
High	16.0	8.1	10.9	7.5

A.10 Distribution (%) of Occupation (1-digit groups) – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages – continued

	Occupation group			
	Managers	*Professionals*	*Sales Workers*	*Technicians & Trade Workers*
CDR	13.3	20.7	9.6	13.4
Gold Standard	13.2	20.7	9.7	13.5
Silver Standard				
Very Low	13.1	21.4	9.4	13.7
Low	12.8	21.2	9.6	13.6
Medium	12.8	21.3	9.5	13.6
High	12.6	22.1	9.3	13.7
Bronze Standard				
Very Low	12.2	22.2	9.4	14.1
Low	11.7	21.9	9.4	13.9
Medium	11.6	21.7	9.5	13.9
High	11.7	22.5	9.2	14.2

A.11 Distribution (%) of Industry of employment (1-digit groups) – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages

	Industry				
	Accommodation & Food Services	*Administrative & Support Services*	*Agriculture, Forestry & Fishing*	*Arts & Recreation Services*	*Construction*
CDR	5.8	2.8	3.6	1.1	5.9
Gold Standard	5.8	2.9	3.7	1.1	5.8
Silver Standard					
Very Low	5.7	2.9	3.3	1.1	5.9
Low	5.7	2.9	3.1	1.1	5.9
Medium	5.7	3.0	3.1	1.1	5.9
High	5.5	2.9	2.7	1.1	5.8
Bronze Standard					
Very Low	5.6	2.8	2.2	1.1	6.0
Low	5.5	2.9	1.7	1.1	5.9
Medium	5.6	3.0	1.8	1.1	5.9
High	5.4	2.9	1.7	1.1	5.9

A.11 Distribution (%) of Industry of employment (1-digit groups) – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages – continued

	Industry				
	Education & Training	*Electricity, Gas & Waste Services*	*Financial & Insurance Services*	*Health Care & Social Assistance*	*Information, Media & Tele-communications*
CDR	7.1	0.8	4.7	10.8	2.2
Gold Standard	7.3	0.8	4.8	11.0	2.1
Silver Standard					
Very Low	7.5	0.8	4.8	11.1	2.1
Low	7.5	0.8	4.9	11.1	2.2
Medium	7.6	0.8	4.9	11.1	2.2
High	7.9	0.8	5.0	11.3	2.2
Bronze Standard					
Very Low	7.8	0.8	5.0	11.3	2.2
Low	7.8	0.8	5.1	11.5	2.3
Medium	7.8	0.8	5.0	11.5	2.3
High	8.3	0.8	5.0	11.7	2.2

A.11 Distribution (%) of Industry of employment (1-digit groups) – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages – continued

	Industry			
	Manufacturing	*Mining*	*Professional, Scientific & Technical Services*	*Public Administration & Safety*
CDR	11.4	0.2	6.3	6.0
Gold Standard	11.5	0.2	6.2	6.1
Silver Standard				
Very Low	11.5	0.2	6.5	6.2
Low	11.6	0.2	6.4	6.2
Medium	11.6	0.2	6.4	6.2
High	11.6	0.2	6.5	6.2
Bronze Standard				
Very Low	11.7	0.2	6.5	6.2
Low	11.9	0.2	6.5	6.2
Medium	11.9	0.2	6.5	6.2
High	11.8	0.2	6.5	6.3

A.11 Distribution (%) of Industry of employment (1-digit groups) – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages – continued

	Industry				
	Rental, Hiring & Real Estate Services	*Retail Trade*	*Transport, Postal & Warehousing*	*Wholesale Trade*	*Other Services*
CDR	1.4	10.8	5.4	4.8	3.6
Gold Standard	1.5	11.1	5.4	4.8	3.7
Silver Standard					
Very Low	1.5	10.9	5.4	4.8	3.8
Low	1.5	11.0	5.4	4.9	3.8
Medium	1.5	11.0	5.4	4.9	3.8
High	1.5	10.8	5.4	4.9	3.8
Bronze Standard					
Very Low	1.5	10.8	5.4	4.9	3.8
Low	1.5	10.9	5.5	5.0	3.8
Medium	1.5	11.0	5.5	5.0	3.8
High	1.5	10.6	5.4	4.9	3.9

A.12 Distribution (%) of Weekly income (for those who answered this question) – CDR, Gold Standard and each level of the Silver Standard and Bronze Standard linkages

	Weekly income			
	Negative	*$0–399*	*$400–999*	*≥ $1,000*
CDR	0.6	46.8	35.7	16.9
Gold Standard	0.6	46.0	36.3	17.1
Silver Standard				
Very Low	0.6	45.0	36.7	17.7
Low	0.6	45.3	36.6	17.5
Medium	0.6	45.0	36.7	17.7
High	0.6	44.2	36.9	18.4
Bronze Standard				
Very Low	0.6	44.3	36.9	18.2
Low	0.5	44.6	37.0	17.9
Medium	0.5	44.6	37.0	17.9
High	0.5	43.5	37.5	18.5

A.13(a) Percentage of linked people whose occupation was Teaching in either 2005 or 2006 or both

	Gold Standard	*Silver Standard – Very Low*	*Bronze Standard – Very Low*
Teacher in 2005 and 2006	71.2	70.7	70.9
Moved out of teaching			
to education related occupation	5.0	4.7	5.1
to education unrelated occupation	3.0	3.3	2.9
to "not in labour force"	4.8	4.4	4.8
Total moved out of teaching	12.8	12.4	12.8
Moved Into teaching			
from education related occupation	5.6	5.7	5.4
from education unrelated occupation	3.5	4.2	4.5
from "not in labour force"	4.3	4.5	4.2
Total moved Into teaching	13.6	14.4	14.1
Missing or NA from one year	2.5	2.6	2.1
Percent of linked pairs with Teacher as occupation in 2005 or 2006	1.6	1.8	1.9
Number of linked pairs with Teacher as occupation in 2005 or 2006	1,157	1,179	1,109

A.13(b) Age ranges of those changing from 'Teacher' in 2005 to 'Not in the labour force' in 2006

	Gold Standard		*Silver Standard – Very Low*		*Bronze Standard – Very Low*	
	Estimate	*95% C.I.*	*Estimate*	*95% C.I.*	*Estimate*	*95% C.I.*
Age						
less than 55 years	74.55	(62.80, 86.30)	71.15	(58.58, 83.72)	73.58	(61.47, 85.69)
55 years and over	25.45	(13.70, 37.20)	28.85	(16.28, 41.42)	26.42	(14.31, 38.53)
Number	50		53		47	

A.13(c) Age ranges of those changing from 'Not in the labour force' in 2005 to 'Teacher' in 2006

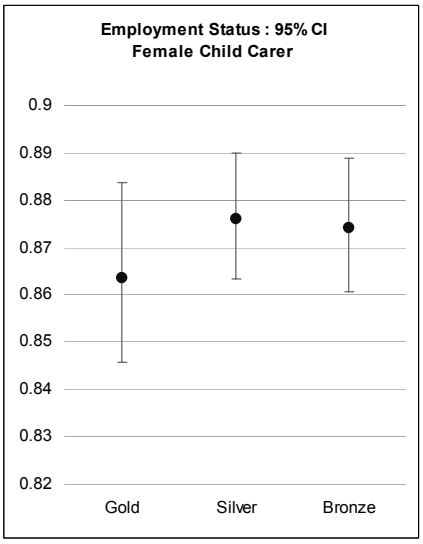
	Gold Standard		*Silver Standard – Very Low*		*Bronze Standard – Very Low*	
	Estimate	*95% C.I.*	*Estimate*	*95% C.I.*	*Estimate*	*95% C.I.*
Age						
less than 25 years	24.00	(11.92, 36.08)	24.53	(12.71, 36.35)	27.66	(14.61, 40.71)
25 years and over	76.00	(63.92, 88.08)	75.47	(63.65, 87.29)	72.34	(59.29, 85.39)
Number	50		53		47	

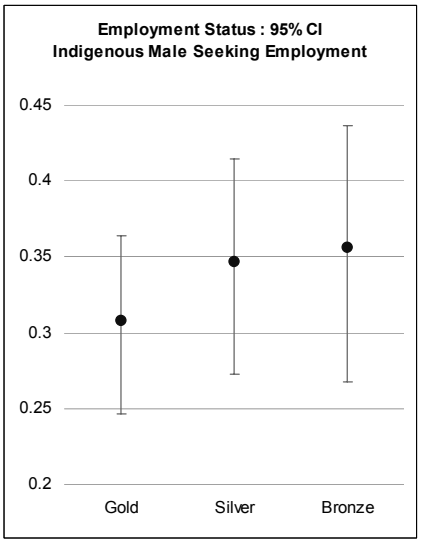
A.14 Employment retention rates among the Indigenous population

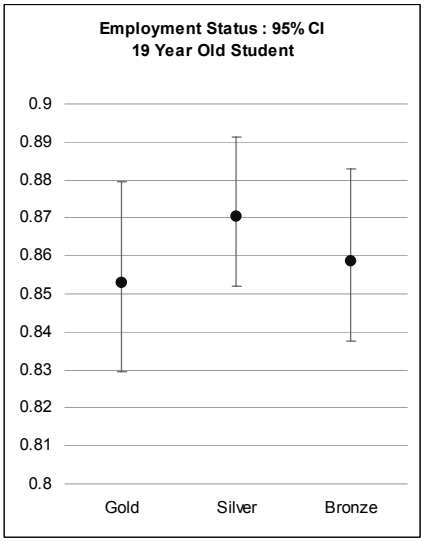
		Status in 2006		
	Status in 2005	*Employed*	*Looking*	*NILF*
Gold Standard				
	Employed	79.8	2.9	17.3
	Looking	34.5	29.3	36.2
	NILF	14.5	6.0	79.5
	Number	987		
Silver Standard – Very Low				
	Employed	86.9	2.9	10.2
	Looking	36.0	34.0	30.0
	NILF	12.2	8.0	79.8
	Number	756		
Bronze Standard – Very Low				
	Employed	88.7	3.1	8.2
	Looking	34.3	34.3	31.4
	NILF	10.9	7.4	81.7
	Number	567		

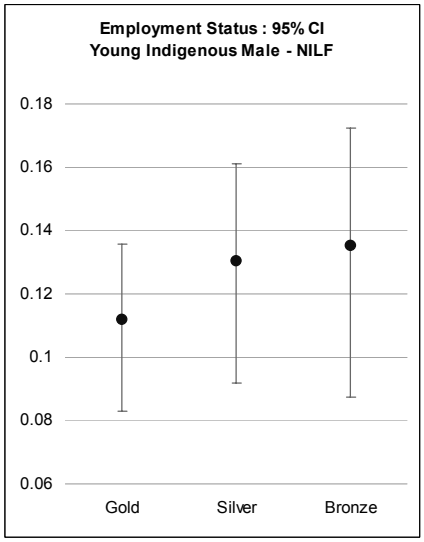
Note: NILF refers to "Not in the Labour Force"

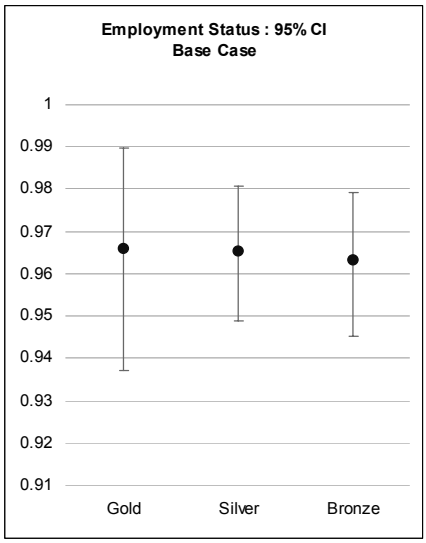
A.15 Test Cases for the odds of employment in 2006 modelled on 2005 variables for Gold Standard, Silver Very Low and Bronze Very Low linked datasets











FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au