1351.0.55.027

Australian Bureau of Statistics

Research Paper

Assessing the Quality of Linking Migrant Settlement Records to Census Data

www.abs.gov.au

New Issue

Research Paper

Assessing the Quality of Linking Migrant Settlement Records to Census Data

Jeffrey Wright, Glenys Bishop

and Tim Ayre

Analytical Services Branch

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 27 AUG 2009

ABS Catalogue no. 1351.0.55.027

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

ABBREVIATIONS

ABS	Australian Bureau of Statistics
CDE	Census Data Enhancement
CDR	Census Dress Rehearsal
CDR_Census	Refers to the dataset formed by linking the CDR to the Census
DIAC	Department of Immigration and Citizenship
RSE	Relative Standard Error
SACC	Standard Australian Classification of Countries
SDB	Settlement Database
SDB_Census	Refers to the dataset formed by linking the SDB to the Census
SDB_SLCD	Refers to the dataset formed by linking the SDB to the SLCD
SLCD	Statistical Longitudinal Census Dataset

ASSESSING THE QUALITY OF LINKING MIGRANT SETTLEMENT RECORDS TO CENSUS DATA

Jeffrey Wright, Glenys Bishop and Tim Ayre
Analytical Services Branch

ABSTRACT

As part of the Australian Bureau of Statistics' Census Data Enhancement project, the Migrants Quality Study was conducted to assess the feasibility of linking the Department of Immigration and Citizenship's Settlement Database (SDB) to the Statistical Longitudinal Census Dataset (SLCD), without the use of name and address as linking variables. This paper provides some background to the Migrants Quality Study, a brief description of the linking process, a thorough evaluation of the quality of the linked data, and then associated discussion about the usefulness of the linked data.

The results from this quality study indicate that linking the SDB to the SLCD is feasible and can produce useful information that no other data source currently provides. However, some quality issues have been identified and need to be thoroughly understood to ensure that the linked data is correctly interpreted and appropriately used.

ACKNOWLEDGEMENTS

This paper was prepared by the Analytical Services Branch, with assistance from the Social Analysis and Reporting Branch and the National Migrant Statistics Unit (NMSU). The NMSU is jointly funded by the Australian Bureau of Statistics and the Department of Immigration and Citizenship.

The Australian Bureau of Statistics acknowledges the assistance provided to this paper by the Department of Immigration and Citizenship.

The authors would like to acknowledge the linking and associated analytical work performed by a team of methodologists, in particular Paul Campbell, Tenniel Guiver, Tom Sullivan, Tetteh Dugbaza and Ben Pang.

The authors would also like to thank David Bartie, Alan Wong, and Peter Rossiter for their valuable assistance and comments in the preparation of this paper.

1. INTRODUCTION

In the lead up to the 2006 Census, the Australian Bureau of Statistics (ABS) initiated the Census Data Enhancement (CDE) project. A key feature of this project is the creation of a Statistical Longitudinal Census Dataset (SLCD), formed by selecting a 5% random sample of records from the 2006 Census, that will subsequently be combined with data from future Censuses. The aim is to bring these datasets together using statistical techniques that do not require name and address. It is envisaged that the SLCD will create a comprehensive picture of Australian society for researchers to investigate issues around improving family well being, employment, health, education outcomes and transitions.

A second aim of the CDE project is to conduct *statistical studies* by bringing the SLCD together with specified non-ABS datasets using statistical techniques. Records will be linked using variables such as date of birth, sex, country of birth and geographic area. At this stage the only planned statistical study is the one to examine conditions of entry and settlement outcomes for migrants, through the linking of the SLCD and the Department of Immigration and Citizenship's (DIAC) Settlement Database (SDB). This is contingent on the outcome of the related *quality study*, discussed below.

The third aim of the CDE project is to conduct *quality studies*, which involves bringing together the full Census data with other datasets. In line with the original proposal and Statement of Intention, the ABS undertook several such quality studies. Two types of quality studies were conducted: studies to improve ABS statistical outputs or to investigate the feasibility and likely quality of linking datasets when name and address are not available. The resulting datasets were destroyed at the completion of the studies. One of these studies (the Migrants Quality Study) involved linking DIAC's SDB to the Census with the purpose of determining whether it would be feasible to conduct the statistical study described above and to identify the likely limitations of the resulting linked data.

The establishment of the CDE project and associated details are presented in ABS information papers *Census Data Enhancement Project: An Update* (ABS, 2006b) and *Enhancing the Population Census: Developing a Longitudinal View* (ABS, 2006a) and the Australian Statistician's *Statement of Intention* (ABS, 2005).

This paper provides a summary of the Migrants Quality Study and makes recommendations about conducting a further statistical study.

2. THE MIGRANTS QUALITY STUDY

The five-yearly Census provides considerable information about migrants living in Australia. However, there are certain questions of great interest to policy makers, researchers and the wider community that Census data alone cannot answer. A key question relates to the relationship between a migrant's visa status and their post-arrival social and economic outcomes. Understanding of settlement outcomes of Australia's migrant and ethnic minority population could be significantly enhanced through analysis of linked Census and SDB data. This would bring together the breadth of the Census data with detailed administrative information from the SDB describing permanent settlers in Australia. By attaching 'visa category' to records on the Census file, variation in family formation, labour market, housing and other socioeconomic outcomes across different migrant groups could be more readily identified and understood. Ultimately this could feed into the future development and evaluation of immigration programs and support services for migrants. Over the longer term, the potential for longitudinal analysis using the Statistical Longitudinal Census Dataset (SLCD) linked to the SDB could lead to improved identification and investigation of the causal factors underlying particular migrant outcomes.

Given these potential benefits, the aims of the Migrants Quality Study were to investigate the feasibility of linking DIAC's SDB with the SLCD and to explore the quality of such linked data. The 2006 Census processing period provided the opportunity to link the full Census with the SDB using two approaches, both with and without name and address as linking variables. Linking with name and address, while not perfect, provides a benchmark for assessing linkage quality when name and address are not available as would be the case when linking with the SLCD. This information has been used to determine whether a statistical study dataset produced from linked SDB_SLCD data would be fit for purpose, and would also aid in the interpretation of results from such a statistical study. It should be noted that all names and addresses collected as part of the Census were destroyed once Census processing was completed.

3. THE DATA

Two versions of the SDB were provided by the Department of Immigration and Citizenship (or its predecessor). Issues with the quality of the data on the first version were discussed with DIAC and these were addressed before the provision of a second version. The first version did not cover all persons who arrived in Australia up to the time of the 2006 Census, but the second version did. In addition, the second version had improved address information, and the categories for the various visa sub-classes had been streamlined. The SDB referred to in this report is the second version provided by DIAC.

The SDB extract used in the Migrants Quality Study covered the period 1 January 2000 to 8 August 2006 (Census night) and contained the records of 861,275 persons who, during that period, were granted visas to live permanently in Australia. Persons on the SDB comprise family, humanitarian, refugee, skilled and State-sponsored migrants. The SDB excludes temporary visa holders and non-visa settlers, such as New Zealanders.

Valuable discussions between ABS and DIAC clarified ABS understanding of the data. The main points are summarised as follows:

1. Data from the SDB are obtained from two sources: a database containing records of persons offshore who apply for and are granted permanent entry visas into Australia (IRIS II) and a database containing records of persons onshore (that is, already in Australia) who apply for and are granted permanent resident visas to remain in Australia (ICSE).

2. A person who applies offshore and is granted a permanent entry visa will be given a grant (approval) number. The person will also be given a visa evidence number when they present their passport to be stamped with the necessary visa to enable travel into Australia to take place. The person's grant (approval) number and date of grant as well as visa evidence number will be added to his or her records on the SDB.

3. A person who applies offshore and is granted an entry visa can still remain on the SDB but will not have an arrival date added to his or her data on the SDB. If the person does not arrive in Australia within 13 months, his or her records will be flagged as 'non-arrival'.

4. In order to avoid matching persons on the SDB who have been granted permanent entry visas but have not arrived in Australia, the key data item to look for is 'arrival date'.

5. For persons who apply onshore for a permanent resident visa to remain in Australia, their arrival date is the date of their last entry into Australia (when they must have entered Australia as students, tourists, temporary workers or provisional spouses etc.). These persons will have an approval number and date of approval in their records on the SDB, but will not necessarily have a visa evidence number unless they present their passport to be stamped.

The 861,275 records on the SDB extract included 53,073 (6.2%) migrants who arrived in Australia after 8 August 2006. Since these persons were not in the country on Census night, they would not have a corresponding record on the 2006 Census file and hence were excluded from the linking process. Further analysis of the SDB extract also identified 1,250 (0.1%) duplicate records. These presumably represent persons who made multiple visa applications and who received multiple acceptances, or possibly offshore applicants who applied for and changed their visa status after arrival in Australia. These duplicate records were also removed from the dataset before linking. Hence after accounting for 'late arrivals' and duplicate records, the number of SDB records available to be linked was 806,952.

The 2006 Census file used for this study consisted of 19,050,146 records, excluding imputed persons and overseas visitors. Imputed persons 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. Overseas visitors are excluded from linking because they are not relevant in this context. These were people who indicated they usually lived in another country and expected to stay in Australia for less than a year. Of the 19,050,146 Census records available for comparison with SDB records, 860,295 indicated they were born overseas with a year of arrival between 2000 and 2006. However, 2,129 were either born on Norfolk Island or inadequately described their country of birth, and approximately 99,000 were identified as people born in New Zealand of whom many would be non-visa settlers and thus not be in the SDB. The figure of 860,295 also included recent arrivals on temporary visas (e.g. students) who would not be in the SDB. A further 212,202 Census records did not state their year of arrival and were either reported as born overseas or had a missing or undefined birthplace. Interestingly, 7,557 of these were aged 5 or less, indicating that some, at least, were recent arrivals.

4. THE LINKING PROCESS

This section provides an overview of the work undertaken by the ABS to create a full Census to SDB linked dataset.

4.1 Linking methodology

The linking methodology used to link the SDB 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 the 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.

This linking methodology can be generalised into the following steps:

- standardisation;
- blocking;
- record pair comparisons; and
- a decision model.

The contents of the two datasets need to be first prepared to allow comparison between the different data sources. Preparation includes a number of steps such as verification, removing inconsistencies and parsing text fields resulting in standardised files and so this step is titled 'standardisation'. This data preparation takes place against a background of determining which variables will be used as linking variables.

Once the data files have been prepared, record pairs, consisting of one record from each file, can be compared to see whether they are likely to be a match, i.e. belong to the same person. However, if the files are large, there may be too many such pairs to conduct the comparison with the resources available. The 'blocking' stage reduces the number of comparisons needed by only comparing record pairs where matches are more likely to be found, e.g. records with the same combination of date of birth, sex, and country of birth.

During the 'comparison' stage, records in corresponding blocks from the two files are compared. Each linking field (variable) for a record pair is compared and the level of agreement is measured by calculation of a field weight. Calculation of field weights depends on obtaining two probabilities for each linking field, the first being the probability that the field values agree if two records belong to the same person, while the second is the probability that the field values agree if the two records belong to different persons.

These are called *m*- and *u*-probabilities, respectively, i.e.

$$m = \Pr\{\text{fields agree} \mid \text{records belong to the same entity}\}; \text{ and}$$
$$u = \Pr\{\text{fields agree} \mid \text{records belong to different entities}\}.$$

These probabilities are estimated using results from similar linking projects and also properties of the two datasets being linked.

Field weights can be modified for a number of circumstances in which there is only partial agreement between linking fields. Some typical comparison options include:

- Exact match (e.g. sex). The fields either agree or they do not, and no adjustment is made to the field weight.
- Exact match (e.g. country of birth) but the weight is modified so that rarer values are given higher weights than more common values when they agree.
- Approximate string comparison (e.g. name). The weight depends on the number of characters that are different, allowing for misspellings, transcriptions of poor handwriting, etc..
- Numerical difference (e.g. date of birth). The weight depends on how far apart values of day or month or year are.
- Geographical difference (e.g. address). Spatial information can be used to calculate distance between fields.

For each record pair comparison, the field weights from each linking field are summed to form an overall record pair comparison weight.

Finally, at the 'decision model' stage, a decision rule determines whether the record pair is linked, not linked or considered further as a possible link. This is done by comparing the record pair comparison weights with cut-off weights. The simplest decision rule is a single cut-off weight when all record pairs with a weight greater than or equal to this weight are assigned as links, and all those pairs with a weight less than the cut-off are assigned as non-links. A more sophisticated decision rule has lower and upper cut-offs. Record pairs with a weight above the upper cut-off are declared links while those with a weight below some lower cut-off are declared non-links. The record pairs with weights between the upper and lower cut-offs cannot automatically be assigned a status and are designated for clerical review.

In clerical review, each record pair is assessed by inspection to resolve match status. Typically, the clerical reviewer is able to identify variations in names and common transcription errors (e.g. 1 and 7) that were not picked up using the comparison options in the 'comparison' stage.

For the case where a record in the first dataset links to multiple records in the second dataset above the upper cut-off, an algorithm is used to optimally assign one record on the first dataset to one record on the second dataset, by maximising the sum of all the record pair comparison weights through alternative assignment choices. This is called one-to-one assignment.

After the 'decision model' stage, the quality of the links is assessed. The methods of evaluation for this quality study are explained in the following sections.

4.2 Implementation in the Migrants Quality Study

Two types of linking were conducted:

- Gold Standard, in which name, address, mesh block and other variables were used for linking; and
- Bronze Standard in which mesh block and other variables were used for linking.

Mesh blocks are micro-level geographical units for statistics and there are in excess of 300,000 mesh blocks covering the whole of Australia. A residential mesh block typically contains 30 to 60 dwellings. A street address can be coded to the appropriate mesh block, but mesh blocks cannot be coded back to a specific street address. This makes them a useful linking variable when street address is not available.

The aim of the Gold Standard linkage was to serve as a benchmark against which the Bronze Standard could be evaluated. The Bronze Standard represents the type of linkage that can be conducted in future statistical studies between the SDB and the SLCD, when name and address are not available. Variables used in the two different standards of linkage are shown in table 4.1.

It should be noted that, in line with the Statement of Intention for the Census Data Enhancement Project, name and address were only available during the Census processing period. At the end of that period, all names and addresses were deleted from linked data files. In addition, the Gold Standard unit record file, which was produced using name and address, was itself destroyed on 31 March 2008. Aggregate confidentialised Gold Standard data have been kept and Bronze Standard unit record linked data are still available.

When performing the Gold Standard linkage, upper and lower cut-offs were used and clerical review was conducted for record pairs with weights between the two cut-offs. The cut-offs were chosen conservatively at first, resulting in very large numbers of record pairs in the range. An acceptance sampling method was implemented, whereby the pairs were ordered according to comparison weight and then divided into batches. A sample of pairs selected from each batch was clerically reviewed. On that basis, the whole batch could be assigned as links, assigned as non-links or sent for full clerical review. This method enabled the cut-offs to be fine-tuned and full clerical review to be conducted optimally on the reduced number of pairs in the middle weight range.

4.1 Variables used to link SDB and Census files for Gold and Bronze Standards

Variable type	*Gold Standard*	*Bronze Standard*
Name information	First name (Actual)	–
	Last name (Actual)	–
	First name (Double metaphone*)	–
	Last name (Double metaphone*)	–
	First name initial (first character of first name)	–
	Last name initials (first two characters of last name)	–
Age-related information	Date of birth	Date of birth
	Age	Age
	Day of birth	Day of birth
	Month of birth	Month of birth
Personal characteristics	Sex	Sex
	Marital status	Marital status
Ethnicity	Country of birth (4-digit)	Country of birth (4-digit)
	Country of birth (2-digit)	Country of birth (2-digit)
	Year of arrival	Year of arrival
	Religion (3-digit)	Religion (3-digit)
	Standardised main language (4-digit)	Standardised main language (4-digit)
Address information	Street number	–
	Street name	–
	Mesh block	Mesh block
	Postcode	Postcode
	Suburb	–

* Double metaphone is a phonetic coding scheme that transforms a string of characters, based on the way the string is pronounced, into a 4-digit code.

To ensure that SDB records were at least compared to their equivalent Census record (belonging to the same person), five linking passes were run, each pass using different blocking variables. This allowed for the situation where two equivalent records might disagree on one blocking variable, but can still be linked in a subsequent pass that uses different blocking variables. Clerical review was performed after each pass, before running the next pass. Approximately five percent of records linked were assigned through full clerical review. Records on each file not linked on one pass were included in the next pass.

Three separate Bronze Standard linkages were conducted, each distinguished by the level of the cut-off weights used; the three linked datasets being Bronze High, Bronze Medium, and Bronze Low. The reason for creating three different Bronze Standard linkages was to enable an assessment of the impact that the level of the cut-off weight has on analysis of the linked data; remembering that the cut-off weight is a measure of the amount of agreement between records necessary to assign them as a link.

For each of the Bronze Standard linkages, only two passes were run (compared to five for Gold Standard) because fewer linking variables were available. For each pass, an acceptance sampling method was implemented to decide on a single cut-off weight (different for each of the three Bronze Standards). All record pairs above the single cut-off weight were linked, and all below it were not linked; those records not linked were sent through to the next pass. For Bronze Standard linkages, clerical review was not used to directly assign individual record pairs as links.

5. 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:

- Comparing the expected number of links between the SDB and Census file to the actual number of links achieved;
- The properties of the SDB records that did not get linked to a Census record;
- Match-link rate and link accuracy of the different Bronze Standard linkages compared with Gold;
- The under- or over-representation of sub-groups in the various linked datasets compared with the Gold Standard; and
- The effects of this under- or over-representation on some typical analyses and models fitted to linked data.

This section provides a summary of the investigations performed.

5.1 Comparing expected number of links to actual number of links

Initially, it is important to consider how many records we might reasonably expect to link. Persons on the SDB file might be missing from the Census file for several reasons:

- they are temporarily out of the country on Census night;
- they are missed by the Census, thus contributing to Census undercount;
- they emigrated from Australia before the Census; or
- they have died since arriving in Australia.

Census undercount estimates indicate that of Australian residents born overseas, the net undercount was approximately 11.9% and this was higher for people from non-English speaking backgrounds (ABS, 2006c). This figure is not directly applicable to the Census data used for linking since it was not the final file to which Census undercount estimates are applied; for example the Census linking file contains no imputed persons. However, it is worth noting that, applying Census net undercount estimates of 11.9% to the SDB, we would expect approximately 96,000 of them not to be on the Census file. This is probably an over-estimate for the reason stated.

It has been estimated that approximately 3,000 Australian residents born overseas who arrived since 2000 would have died before the Census. This is based on death rates available from the publication *Deaths, Australia* (ABS, 2007b).

On the 2006 Census night, 345,200 people were temporarily out of the country (see *Australian Demographic Statistics*, ABS, 2007a). A conservative estimate is that 30% of these were born overseas and about 15% of those were recent arrivals. Thus, using a pro rata method, we might expect 15,000 recent arrivals to be out of the country on Census night. There would have been some emigrants as well.

Applying all these factors very approximately indicates that we would expect to link somewhere in the vicinity of 700,000 SDB records to the 2006 Census.

Table 5.1 shows the actual number of links achieved for the Gold and Bronze Standards, and also the total number of SDB records available for linking.

5.1 Number of SDB records available for linking and the numbers linked for Gold Standard and each level of Bronze Standard

	Number of records linked			
SDB records	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
806,952	511,066	340,110	429,439	529,695

Note: Bronze High, Medium, and Low refer to the three Bronze Standard linkages; High, Medium, and Low indicate the level of cut-off weights used for each linkage.

Table 5.1 shows that substantially fewer than the expected number of links were made in the Gold Standard. Bronze Standard files will generally have fewer links because not as many linking variables are available. As the cut-off is lowered, more links are made. The extra links observed in the Bronze Low file, compared to Gold, may be due to two possibilities:

- These records should have been linked in the Gold Standard but did not meet the stringent criteria; for instance, name might have been missing from one or both records and if the weight was in the clerical review range, the link might not have been confirmed; or
- These records are not true links but belong to people with similar characteristics.

5.2 SDB records that were not linked

The main reasons for not linking records on the SDB to Census are:

- the corresponding Census record does not exist (this was discussed in 5.1); and
- the quality of data on either the SDB record or the corresponding Census record is too poor to allow a link to be made.

Data quality can be affected by respondents not completing key questions or making errors in the information provided. This issue affects both datasets. Table 5.2 contains details of data items missing from the SDB for records not linked in the Gold Standard.

5.2 Key linking variables missing on SDB records not linked in the Gold Standard

Linking variable	*Number of records missing variable information*	*Percentage of 295,886 unlinked SDB records*	*Percentage of 806,952 SDB records*
First name	4,096	1.4	0.5
Sex	29	0	0
Country of birth	10,685	3.6	1.3
Language (a)	125,739	42.5	15.6
Marital status (age ≥ 18 years)	43,257	18.0 (b)	6.7 (c)
Year of arrival	1,345	0.5	0.2
Mesh block	108,873	36.8	13.5
Postcode	54,779	18.5	6.8
Street number	103,906	35.1	12.9
Street name	99,632	33.7	12.3
Suburb	95,806	32.4	11.9

(a) An entry for language exists for all records on the SDB; however 320,882 records had codes that were 'inadequately described' and so could not be standardised to the Census. They were therefore treated as 'missing'. A total of 125,739 of these records could not be linked on the Gold Standard.

(b) Based on 240,829 unlinked SDB records of persons aged 18 years and over.

(c) Based on 648,052 SDB records of persons aged 18 years and over.

Since address details are required to allocate a mesh block code, it is expected that most of the records missing street number and street name are also missing mesh block. Those missing language and country of birth may or may not overlap with these. Table 5.2 suggests that approximately 100,000 SDB records could not be linked because of inadequate address information, and address details are key linking variables.

From discussions with DIAC, another issue to consider is whether or not address information on the SDB is up to date. Even if the address fields on the SDB are completed, they may not correspond to the same address as reported on the 2006 Census. In this case, the comparison of address information would yield a negative comparison weight, and thus reduce the likelihood that the records would be linked. From a linking perspective, outdated addresses are more problematic than missing addresses when trying to establish agreement between records. The extent of this problem of outdated address information on the SDB is somewhat unknown, but it is expected to be one of the major reasons for the large number of unlinked SDB records (for Gold and Bronze Standards).

Other factors associated with the SDB that could influence linking are considered below.

1. Language was poorly reported, with about 42 percent of unlinked records reporting inadequately described language details that could not be standardised to an equivalent code used in the Census.

2. The impact of missing variables was in some cases compounded by the fact that for some records, more than one linking variable was missing. For example, one-third of unlinked records had four or more linking variables missing, while one-quarter had five or more variables missing.

3. The linking process also highlighted factors specific to particular migrant groups that impacted on linking outcomes. For example, of the approximately 7,200 migrant records on the SDB missing 'first name', 44% were born in India.

4. It was also noted that disproportionately large groups of migrants from some countries report the same date of birth. These dates appear to be allocated administratively rather than actual birth dates. Provided that those people consistently use the same birth date (on Census too) this will not prevent them being linked. However, for these people, date of birth does not have the same distinguishing power among persons from the same country that it would usually have for persons from other countries.

As mentioned above, poor data quality on the Census form could also cause difficulty in linking. Table 5.3 shows that about 14% (or 150,000) of relevant Census records have three or more missing linking variables for the Gold Standard linkage. In this instance, relevant Census records refer to people who indicated they were born overseas with a year of arrival between 2000 and 2006, or were people who did not state their year of arrival and were either reported as born overseas or had a missing or undefined birthplace. The number of these types of records is 1,072,497. Note that although Census records are subset here to calculate indicative missing rates for migrants, all 19,050,146 Census records were considered in the linking.

5.3 Number of Gold Standard linking variables missing on relevant Census records (1,072,497 records as defined in paragraph above)

Number of linking variables with missing values	*Number of relevant Census records*	*Percentage of 1,072,497 relevant Census records*
0	550,779	51.4
1	308,088	28.7
2	63,742	5.9
3	51,368	4.8
4	28,533	2.7
5 or more	69,987	6.5

Table 5.3 shows that some SDB records will remain unlinked because the quality of information on the equivalent Census record is insufficient to establish enough agreement to assign a link.

In summary, consider the 295,886 unlinked SDB records in the Gold Standard linkage. The above discussion has established that approximately 100,000 records were unlinked because no equivalent Census record existed, 100,000 remained unlinked because of missing address information, a potentially large number remained unlinked because of outdated address information on the SDB, and some remained unlinked because of poor reporting or missing values on Census records.

The same issues relate to the Bronze Standard linkages, however the issues relating to quality of address information (coded to mesh block for Bronze Standard) are even more significant because without the use of name as a linking variable, mesh block is relied upon even more to establish agreement between records.

5.3 Match-link rate and link accuracy

Matches are defined as record pairs in which the two records relate to the same person. If we consider the Gold Standard links as matches, then we can use them as a benchmark for the Bronze Standard linkages. We are interested in the proportion of links in a given Bronze dataset that are matches (the link accuracy), and the proportion of possible matches that are actually linked in the given Bronze dataset (the match-link rate). Match-link rate and link accuracy were calculated for each linkage level of the Bronze Standard by comparing them with the Gold Standard as shown in table 5.4.

5.4 Method of calculating Match-link rate and Link accuracy

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

For Bronze Low, there were 529,695 links in total and 424,853 of these corresponded to Gold links, ie 424,853 true links. In the Gold Standard, there were 511,066 links, which are designated as total matches in table 5.4. Thus the values for match-link rate and link accuracy for Bronze Low are calculated as follows:

$$\text{Match-link rate} = \frac{\text{True links}}{\text{Total matches}} = \frac{424,853}{511,066} = 83.1\%$$

$$\text{Link accuracy} = \frac{\text{True links}}{\text{Total links}} = \frac{424,853}{529,695} = 80.2\%$$

The results for the three Bronze Standard levels are presented in figure 5.5. The figure shows that as the cut-off is lowered, the match-link rate increases, while the link accuracy decreases. The question arises as to which is more important, high accuracy or high match-link rate. This is addressed in Sections 5.4 and 5.5.

5.5 Match-link rate and Link accuracy for Bronze Standard linkages between Migrants and Census

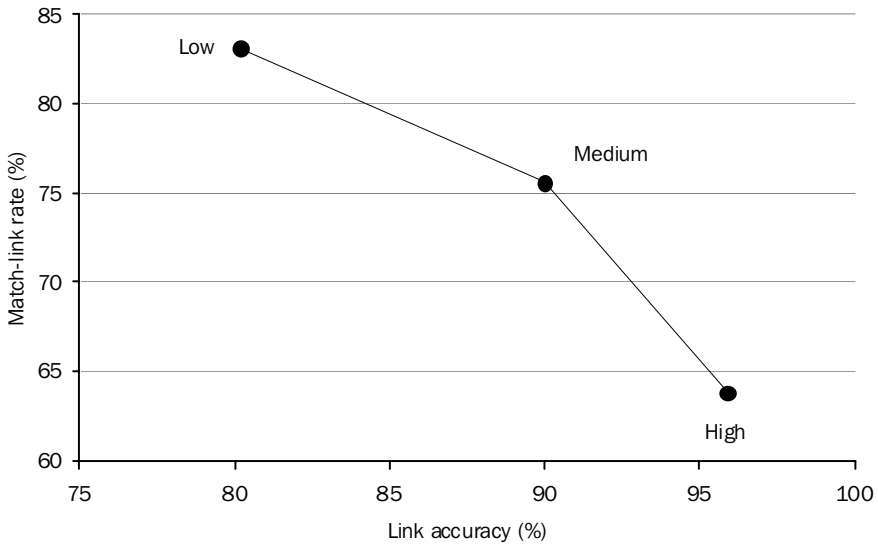
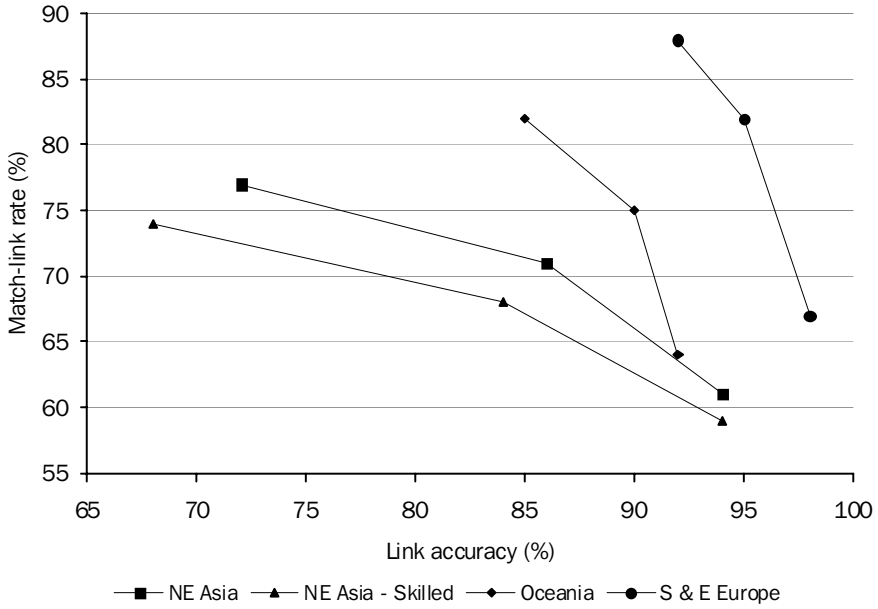


Figure 5.6 shows match-link rates and link accuracy for selected regions of birth; the regions being defined by the 1-digit level of the Standard Australian Classification of Countries (SACC). The 1-digit level of SACC classifies countries into nine broad regions (only three of them are shown in figure 5.6). North East Asia generally has the lowest quality linkage of all the broad regions, and Southern & Eastern Europe has the highest. Also shown in figure 5.6, are migrants on skilled visas born in North East Asia, which have the lowest overall quality of all the significant broad region by visa category classifications.

5.6 Match-link rate and Link accuracy, by Country of birth (selected groups)



5.4 Under- and over-representation of sub-groups

Analysis of various migrant characteristics from the SDB extract and the Gold and Bronze Standard linked files was undertaken and the relative proportions compared. The analysis indicated that no major subpopulations were missed in great numbers in the linking process; however, some groups do seem to be more difficult to link than others, resulting in some degree of under-representation on the linked files. It was found that rates of under-representation on the Gold Standard file were highest for migrants for the following groups:

- Migrants on skilled visas, particularly those in any of the following categories:
    - skilled migrants aged 25–34 years;
    - skilled migrants born in East Asia; or
    - skilled migrants who received 'offshore' visas;
- Migrants who reported speaking English 'very well';
- Migrants who had never been married; and
- Migrants who reported having no religion.

Further details on these findings are presented in tables 5.7 (a), (b), (c), (d), (e) and (f), which also include information on the Bronze Standards.

5.7(a) Relative frequencies (%) in each Visa category, for Gold and Bronze Standard linked data compared with the SDB

Visa category	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Skilled	52.0	48.2	45.8	46.8	48.9
Family	37.5	39.6	40.9	40.3	39.2
Humanitarian	9.5	11.1	12.3	11.9	10.9
Other	0.9	1.0	1.0	1.0	1.1

5.7(b) Relative frequencies (%) in each English proficiency category, for Gold and Bronze Standard linked data compared with the SDB

English proficiency	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Very well	62.7	59.1	56.0	56.5	59.2
Well	7.5	8.0	8.5	8.5	8.1
Not well	17.7	19.1	20.4	20.3	19.2
Not at all	12.1	13.8	15.1	14.7	13.4

5.7(c) Relative frequencies (%) in each Marital status category, for Gold and Bronze Standard linked data compared with the SDB

Marital status	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Married	49.2	52.2	52.5	52.6	50.6
Never married	39.9	36.9	36.7	36.8	38.7
Divorced /separated	1.7	1.6	1.6	1.5	1.6
Widowed	1.0	1.1	1.1	1.1	1.0
Engaged /de facto	8.2	8.2	8.2	7.9	8.1

5.7(d) Relative frequencies (%) of Skilled migrants in each Age group, for Gold and Bronze Standard linked data compared with the SDB

Age group	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
0–14 years	20.0	21.2	22.0	20.7	20.5
15–24 years	12.6	12.9	13.5	14.0	14.2
25–34 years	31.1	28.6	27.0	28.0	29.8
35–44 years	25.0	26.4	26.2	26.3	25.4
45–54 years	9.4	9.3	9.7	9.4	8.7
55–64 years	1.6	1.4	1.5	1.4	1.2
65+years	0.3	0.2	0.2	0.2	0.2

5.7(e) Relative frequencies (%) of Skilled migrants, by Region of birth, for Gold and Bronze Standard linked data compared with the SDB

Region of birth	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Oceania	3.2	3.8	3.9	3.7	3.3
NW Europe	23.5	25.0	23.4	24.3	24.6
South & East Europe	2.3	2.6	2.4	2.4	2.2
North Africa & Mid East	2.2	2.1	2.1	2.1	2.0
SE Asia	17.5	14.5	15.5	15.2	15.2
NE Asia	18.7	17.4	17.0	17.3	17.9
South & Central Asia	17.6	18.7	19.6	19.3	19.6
Americas	2.9	2.8	2.7	2.7	2.5
Sub-Saharan Africa	12.1	13.3	13.4	13.1	12.7

5.7(f) Relative frequencies (%) of Skilled migrants, by Place of visa application, for Gold and Bronze Standard linked data compared with the SDB

Place of visa application	*SDB*	*Gold*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Onshore	29.7	32.1	31.9	31.2	30.9
Offshore	70.3	67.9	68.1	68.8	69.1

In general, skilled visa holders are under-represented on all linked files while family and humanitarian visa holders tend to be over-represented. The extent of over- or under-representation is less marked for Bronze Low than for Bronze High and Bronze Medium. Bronze Low resembles the Gold Standard more closely than the other levels. This can probably be explained by the fact that records that are harder to link, because of missing linking variables, are more likely to be linked when the cut-off is lowered.

Considering tables 5.7 (a), (b), (c), (d), (e) and (f), the observed differences between the SDB and the Gold Standard are to some extent reflective of broader trends in statistical monitoring, where younger, more mobile groups within the community tend to be more difficult to enumerate, and hence more difficult to link.

For example, skilled migrants are likely to be more self-sufficient than other migrant groups, particularly in cases where they have high levels of proficiency in English and few family commitments. Such self sufficiency may reduce the extent to which skilled migrants seek government support and interact with administrative systems that are used to update address information on the SDB. In addition, some migrants, depending on their visa arrangements, may not be immediately eligible for government support upon their initial arrival in Australia.

These factors, in combination with a higher propensity for residential and occupational mobility (as with other younger populations), may serve to reduce both the volume and quality of available data on the SDB and Census files. This has a subsequent impact on the quality of the linked data. Similar factors are likely to apply in the case of offshore applicants, for whom post-arrival address details, in particular, may not be as reliable as for those who apply onshore. Reasons for higher rates of under-representation amongst skilled migrants from East Asia are less obvious, but may again reflect the mobility and administrative data issues discussed above.

In contrast to those groups under-represented on the Gold Standard, some migrant populations such as humanitarian migrants are somewhat over-represented. To a large extent this can be attributed to their greater degree of engagement with administrative systems such as Medicare.

5.5 Some typical analyses and the impact on conclusions from using different linkage standards

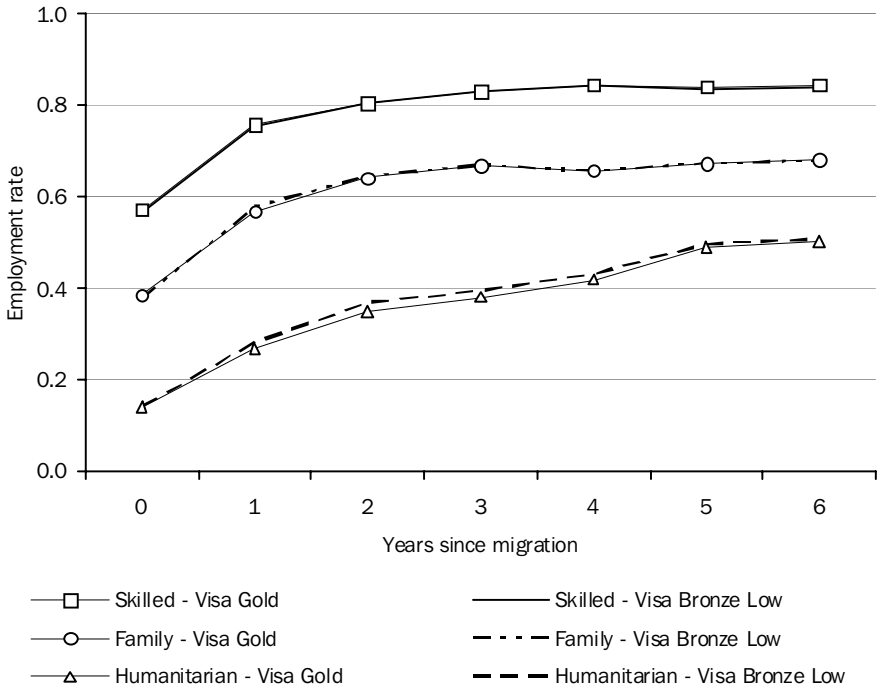
DIAC provided the ABS with a number of issues for investigation using linked data. One such issue is the level of employment for migrants with respect to arrival cohorts, and whether it varies for different visa categories. This issue has been investigated using the Gold and Bronze Standard linked data, to see what effect the different linkage standards have on the conclusions drawn from the analysis.

Preliminary analysis of the three Bronze Standards indicated that Bronze Low produces the most similar conclusions to Gold Standard, hence analysis presented in this report was restricted to comparing Bronze Low to Gold Standard.

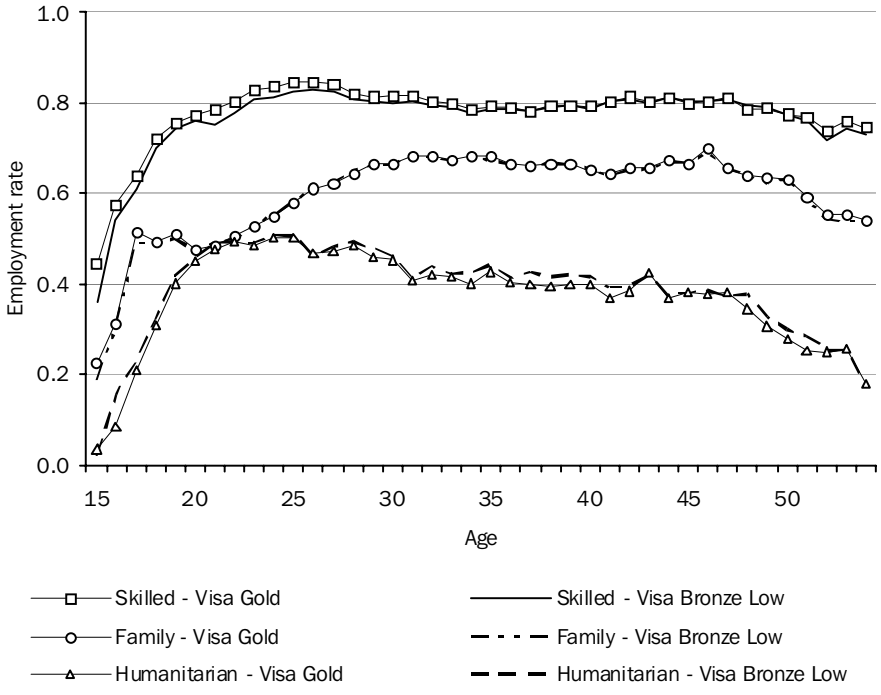
Figure 5.8 shows the relationship between employment rate and years since migration for different visa categories. Bronze Low tracks Gold very well for skilled visa holders and very closely beyond three years for family visa holders. However, for humanitarian visa holders, the Bronze Low curve is systematically above the Gold curve, meaning that the observed relationship is slightly different for the two linked datasets.

Figure 5.9 shows the relationship between employment rate and a person's age for different visa categories. The Bronze Low curve for skilled visa holders tracks below the Gold curve, particularly for younger ages, meaning that there is a difference in the observed relationship. The Bronze Low and Gold curves for family visa holders are very similar. For humanitarian visa holders, the Bronze Low curve tracks above the Gold curve for most ages.

5.8 Proportion of each Visa category employed, by Years since migration: Gold vs Bronze Low



5.9 Proportion of each Visa category employed, by Age: Gold vs Bronze Low



To understand the importance of differences between Bronze Low and Gold Standards, consider 40-year-old humanitarian visa holders. Looking at figure 5.9, the difference in employment rates between Gold and Bronze Low is about two percentage points. By way of comparison, an equivalent estimate obtained from a sample survey such as the *Labour Force Status and Other Characteristics of Recent Migrants Survey* (ABS, 2007c) would have a confidence interval substantially larger than the two percentage point difference observed above. This means that the error due to using the Bronze Low Standard (compared to Gold) is relatively small when compared to the sampling error.

Generally speaking, it appears that the relationships observed in figures 5.8 and 5.9 are quite similar for Gold and Bronze Low Standards. That is, Bronze Low would still allow some useful analysis of employment rate by other characteristics.

To further investigate these relationships and identify differences between the Gold and Bronze Standards, some work has been done to fit a model of employment status (Employed, Unemployed or Not in the Labour Force) using visa category and several Census variables as explanatory variables. The explanatory variables included in the model are listed below:

- Age;
- Sex;
- Registered marital status;
- Proficiency in spoken English;
- Year of arrival;
- Highest level of school completed;
- Level of tertiary education;
- Full/Part-time student status;
- Child caring responsibilities;
- Place of usual residence (urban or regional);
- Visa category (skilled, family, humanitarian, other) (from SDB); and
- Place of visa application (onshore or offshore) (from SDB).

The model was fitted to the Gold Standard, and then a similar form of model using the same explanatory variables was fitted to the Bronze Low and Bronze Medium Standards to enable comparison of the model results. The parameter estimates from the models indicate that Bronze Low is more similar to the Gold Standard than is the Bronze Medium.

A simple summary statistic of such an evaluation is the 'deviance for coefficients' as defined by Chipperfield (2009). This deviance statistic can be interpreted as the average difference between a Bronze regression coefficient and its corresponding Gold coefficient, where the difference is measured in terms of the number of standard errors of the Gold coefficient. For Bronze Low, the deviance statistic was 1.89, compared to 6.24 for Bronze Medium. These results indicate that the Bronze Low Standard's coefficients are closer to the corresponding Gold Standard coefficients. Although a model was not fitted to the Bronze High Standard (due to time constraints), other linkage studies indicate that Bronze High would have a higher deviance statistic than for Bronze Medium. This is because the Bronze High Standard is less representative of the entire population (see Section 5.4).

Some of the practical effects of using the different linkage standards can be seen by considering model predictions, shown in table 5.10, for the following example.

Example person

Male, 35–44 years old, married, humanitarian visa, applied offshore, does not speak English well, arrived in Australia six years ago (from Census 'year of arrival' variable), has completed school above year 10 or equivalent but has no post-school qualifications, not currently studying, and not living in a major city.

5.10 Predicted probabilities of Employment status, by different Linkage standards, for the example person described above

Linkage standard	*Pr(employed)*	*Pr(unemployed)*	*Pr(not in the labour force)*
Gold	0.6404	0.0695	0.2901
Bronze Low	0.6411	0.0748	0.2841
Bronze Medium	0.4736	0.0639	0.4625

According to the model fitted to the Gold Standard, the probability of the example person being employed was 0.6404; the probability of being unemployed was 0.0695; and the probability of being not in the labour force was 0.2901. The model fitted to Bronze Low Standard yielded similar results to the Gold Standard model, but the model fitted to Bronze Medium Standard produced quite different predicted probabilities for employed and not in the labour force.

From the brief model investigations described above, Bronze Low appears to produce results most similar to Gold.

5.6 Evaluation summary

In summary, about 36.7% of the SDB could not be linked to the Census data when name, address, mesh block and other variables were used for linking (Gold Standard). About one third of these can be explained by known reasons that a Census record would not exist. A further one third is probably due to missing and incomplete responses on the SDB, particularly address items and language. The remaining third are most likely due to outdated address information on the SDB and poor quality responses for some Census records.

An issue to consider is the diligence that recent migrants apply to the completion of forms. Forms may not be completed thoroughly because questions are not understood.

The under- and over-representation of particular groups mean that care should be taken when interpreting analyses about those groups. However, it may be possible to overcome this problem by calibrating the linked data. This is discussed in the next section.

We have found from extensive analysis of other linked data that missed links usually cause more problems with drawing conclusions from linked data than do incorrect links. Chipperfield (2009) discusses some aspects of this. In the case of the linked SDB_Census data, we see similar trends with Bronze Low more closely resembling the SDB than Bronze High or Bronze Medium. Our recommendation is that for future statistical research, the Low cut-off should be used.

It is also useful to see whether the conclusions we have drawn from the linked SDB_Census data apply to the subset of SDB records linked to the SLCD (5% of Census). This is also discussed in the next section.

6. OTHER INVESTIGATIONS

6.1 Calibration

It may be desirable to have estimates of numbers of migrants in various categories sum to the known total number of migrants. For instance, in a future statistical study, DIAC may wish to make statements about the total number of migrants, who arrived between 1 January 2000 and 8 August 2006, in each of several different income categories. By calibrating the linked unit record file to known subpopulation totals in the SDB, it would be possible to obtain such numbers.

It may be possible to overcome the differential linking rates of certain population sub-groups in the Bronze Low file by calibration if the calibration variables explain differences in linking rates.

A calibration method was tested as part of this quality study. The Bronze Low linked dataset was calibrated to SDB totals for sex, age, visa category, year of arrival, country of birth, whether the visa had been granted onshore, and whether a person was the main applicant for a visa. This is very fine calibration and unfortunately does not leave any variables in the SDB that can be used to assess whether the calibration has caused any bias in the estimates nor whether calibration has overcome any differential in the linking rates.

We can say, however, that calibration did not make much difference to the distributions of income and of English proficiency.

6.2 The SLCD subset of the linked data

Using the SDB_Census linked dataset (formed using 100% of Census), a subset was formed by selecting only those links with a Census record in the SLCD (5% of Census); this dataset was named the SDB_SLCD linked file. A SDB_SLCD linked file was created for each of the Gold and three Bronze Standards. Using these datasets, match-link rate and link accuracy were re-calculated for the links. Figure 6.1 shows that linking of the SDB to the SLCD subset performed a little better than linkages to the whole Census dataset, in terms of match-link rate and link accuracy. The values for the SDB linked to the full Census in figure 6.1 are the same as those displayed in figure 5.5.

6.1 Comparison of Match-link rate and Link accuracy for SDB linked to full Census and SLCD subset

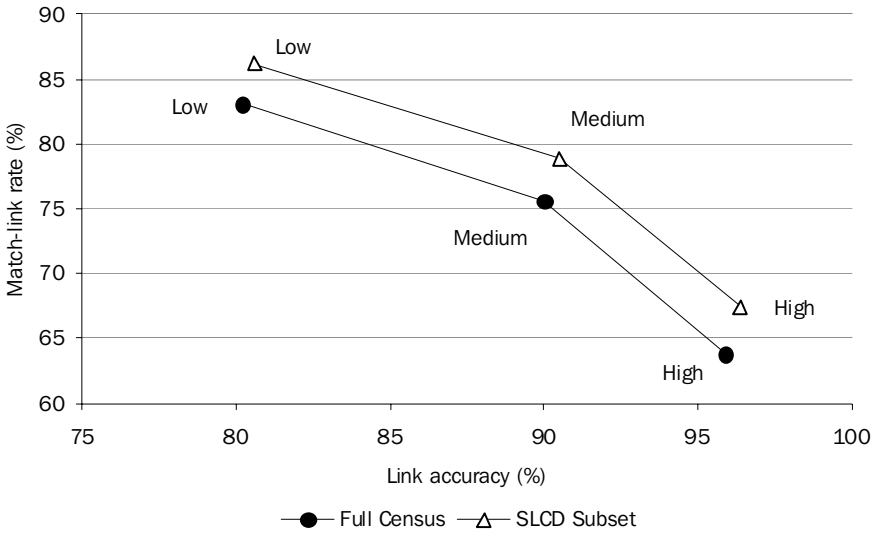
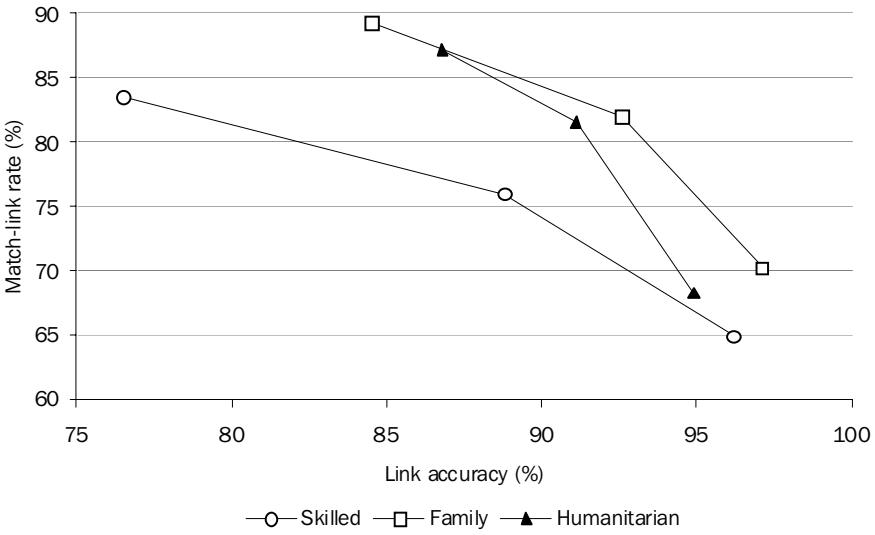


Figure 6.2 shows the link accuracy and the match-link rates for the three main visa categories in the SLCD subset. Overall, based on these measures, skilled migrants appear to have the lowest quality linkage, and family migrants have the highest quality linkage. Across the three Bronze Standards, humanitarian migrants have the lowest variation in link accuracy, and skilled migrants have the highest. Match-link rate varies by about the same amount for all three visa categories.

6.2 Match-link rate and Link accuracy, by Visa category (SLCD subset)



In general, the SLCD subset of the linked data has tended to retain the under- and over-representation of sub-groups that exists on the full Census linked datasets (discussed in Section 5.4). An exception to this is that migrants on humanitarian visas are less over-represented on the SLCD subsets of Bronze High and Bronze Medium than they are on the corresponding datasets linked to the full Census and are slightly under-represented on the SLCD subset of Bronze Low. Correspondingly, migrants on skilled visas appear to be less under-represented. Details of the distribution of visa category on Bronze Standard subsets are shown in table 6.3.

6.3 Relative frequencies (%) in each Visa category, for the Bronze Standard SLCD subset compared with the SDB

Visa category	*SDB*	*Bronze High*	*Bronze Medium*	*Bronze Low*
Skilled	52.0	47.6	48.3	50.2
Family	37.5	41.7	40.9	39.7
Humanitarian	9.5	9.9	9.9	9.1
Other	0.9	0.9	0.9	1.0

Table 6.4 shows the distribution of region of birth for skilled migrants on the SDB and on the SLCD subset of Bronze Standard linked data. The under- and over-representation of groups are very similar to those observed for the Bronze Standard linkage with the full Census (table 5.7(e)).

6.4 Relative frequencies (%) of Skilled migrants, by Region of birth, for Bronze Standard SLCD subset compared with the SDB.

Region of birth	*SDB*	*Bronze high*	*Bronze medium*	*Bronze low*
Oceania	3.2	3.9	3.6	3.2
NW Europe	23.5	23.5	24.5	24.7
South & East Europe	2.3	2.6	2.6	2.3
North Africa & Mid East	2.2	2.3	2.2	2.1
SE Asia	17.5	15.6	15.3	15.3
NE Asia	18.7	16.9	17.4	18.2
South & Central Asia	17.6	19.4	19.0	19.4
Americas	2.9	2.4	2.5	2.5
Sub-Saharan Africa	12.1	13.5	13.0	12.4

6.3 Impact of sampling error on reliability of estimates

Since the SLCD is a 5% sample of the Census population, there will be sampling error associated with estimates from the SDB_SLCD linked file. Sampling error is the error in the estimate caused by selecting a sample rather than taking a complete enumeration of the population. It is caused by the variability in responses for different members of the population. Generally speaking, as the sample size increases, the amount of sampling error decreases. Using a 5% sample, there will be some subpopulations for which there will be too few observations in the sample to obtain reliable estimates (the sampling error is too large). The point of interest is at what subpopulation size the sampling error becomes so great as to make an estimate unreliable. For example, does the 5% sample contain enough observations to produce reliable estimates of employment rate by years since migration for migrants on humanitarian visas at the state level, or could reliable estimates only be produced at the national level for such a breakdown?

The level at which reliable estimates could be produced will largely dictate whether the SDB_SLCD linked file can add extra statistical information to what already exists from other data sources. A simple comparison of sample sizes between the SDB_SLCD linked file and other migrant data sources (that contain similar analysis variables) can give an indication of whether the SDB_SLCD linked file will be able to produce statistics that other data sources cannot. Tables 6.5(a), (b) and (c) show the number of records in the SDB_SLCD linked file by various characteristics.

Tables 6.5(a), (b) and (c) show that the SDB_SLCD linked file contains 28,262 records, with 23,666 records being for people aged at least 15 years.

6.5(a) Number of records in SDB_SLCD linked file, by Age and Visa category

	Visa category				
Age	*Skilled*	*Family*	*Humanitarian*	*Other*	*Total*
Less than 15 years old	2,844	926	761	25	4,596
≥ 15 years old	11,311	10,300	1,796	259	23,666
Total	14,195	11,226	2,557	284	28,262

6.5(b) Number of records in SDB_SLCD linked file, by Application details (main applicant or dependent, offshore or onshore application) and Visa category, for people aged at least 15 years

	Visa category				
Application details	*Skilled*	*Family*	*Humanitarian*	*Other*	*Total*
Main applicant / Applied offshore	3,906	5,261	615	22	9,804
Main applicant / Applied onshore	2,611	4,129	275	124	7,139
Not main applicant / Applied offshore	3,673	665	825	40	5,203
Not main applicant / Applied onshore	1,121	245	81	73	1,520
Total	11,311	10,300	1,796	259	23,666

6.5(c) Number of records in SDB_SLCD linked file, by State/Territory and Visa category, for people aged at least 15 years

	Visa category				
State	*Skilled*	*Family*	*Humanitarian*	*Other*	*Total*
New South Wales	4,008	4,420	746	141	9,315
Victoria	3,091	2,711	533	76	6,411
Queensland	1,630	1,453	147	18	3,248
South Australia	648	452	145	5	1,250
Western Australia	1,613	991	164	15	2,783
Tasmania	84	75	31	*	*
Northern Territory	56	66	17	*	*
ACT & Other Territories	181	132	13	*	*
Total	11,311	10,300	1,796	259	23,666

* cell totals suppressed for confidentiality reasons

Other data sources that contain similar information to the SDB_SLCD linked file include:

- the ABS *Labour Force Status and Other Characteristics of Recent Migrants Survey* (ABS, 2007c);
- the ABS *General Social Survey* (ABS, 2006d); and
- DIAC's *Longitudinal Survey of Immigrants to Australia* (LSIA) (DIAC, 2009).

The *Labour Force Status and Other Characteristics of Recent Migrants Survey* conducted in November 2007 included migrants who arrived in Australia after 1997, were aged 15 years and over on arrival, and who had obtained permanent Australian resident status. In addition, the survey included people who were temporary residents of Australia for 12 months or more, and excluded those born in New Zealand, those holding New Zealand citizenship and those who held Australian citizenship prior to their arrival in Australia. The response rate to the survey was 95%, yielding a sample of 2,530.

The *General Social Survey* conducted from March to July 2006 included all people (not just migrants) aged 18 years and over, resident in private dwellings, throughout non-remote areas of Australia. The response rate was 86.5% yielding a sample of 13,375. Note, however, that only 3,428 people in the sample were born overseas, and when further restricted to people with year of arrival between 2000 and 2006 (inclusive), the number drops to 433.

The latest version of DIAC's *Longitudinal Survey of Immigrants to Australia* (LSIA 3) included primary applicants from the SDB, at least 18 years of age, with an identifiable country of birth but excluded migrants with special eligibility visas, and New Zealand citizens. The response rate for wave 1 of the survey (6 months after arrival) was approximately 48% and yielded a sample of 9,900.

In terms of sample size, the SDB_SLCD linked file has the largest sample by far. This means that overall the SDB_SLCD linked file would have less sampling error and thus be able to produce reliable estimates at finer levels compared to the other similar data sources currently available. Some analyses would still be restricted by sampling error, but less so than the other data sources. However, the linked dataset may have more bias than the two ABS surveys, caused by the linking process. The longitudinal survey (LSIA) is likely to have a substantial non-response bias.

To demonstrate the practical effect that sampling error could have on analysis of the SDB_SLCD linked file, sampling errors for a typical analysis were calculated. The errors calculated were for estimates of employment status rates (unemployed, employed, not in the labour force) by years since migration (single years) by visa category (skilled, family, humanitarian) by state. Results of this analysis are presented in table 6.6. The measurement of sampling error used in table 6.6 is the Relative Standard Error (RSE). Estimates with RSEs less than 25% are considered sufficiently reliable for most purposes. Estimates with RSEs between 25% and 50% should be used with caution, and estimates with RSEs greater than 50% are considered too unreliable for general use.

6.6 Relative standard errors for estimates of Employment status rates, by Years since migration (single years), by Visa category, by State

		Estimates by Years since migration			
Visa category	*Employment status*	*National*	*Large States (NSW, Vic.)*	*Medium States (Qld, SA, WA)*	*Smaller States (Tas., NT, ACT)*
Skilled /Family	Employed & NILF				
	Unemployed				
Humanitarian	Employed & NILF				
	Unemployed				

Legend:

RSE	*Advice*	*Colour code*
<25%	Reliable	
25%– 50%	Caution	
> 50%	Unreliable	

* The column titles (Large States, Medium States and Smaller States) are based on the number of migrants in each State /Territory

* NILF = Not in the Labour Force

In table 6.6, the skilled and family visa categories and also estimates of employed and not in the labour force have been included together because these subpopulation sizes are similar and thus have similar RSEs.

Table 6.6 shows that for smaller subpopulations, such as unemployed migrants on humanitarian visas, the estimates of years since migration (in single years) become unreliable when produced for medium and smaller states. However, for larger subpopulations, such as employed migrants on skilled visas, it is still possible to produce reliable estimates for medium states.

It should be noted that table 6.6 presents RSEs for estimates by years since migration (single years). As estimates by these categories become unreliable, it would be possible to collapse categories and produce more reliable estimates.

With consideration of the size of the subpopulations in the analysis above, this analysis indicates that the SDB_SLCD linked file should be able to produce some reliable estimates at the state level (at least for larger states) for some typical analyses.

Although the sampling error associated with the SDB_SLCD linked file would restrict its usefulness for some fine level analyses, in light of other existing migrant data sources, the sampling error does not severely restrict the usefulness of the SDB_SLCD linked file in terms of the information gaps it could fill.

6.4 Future linking with SLCD

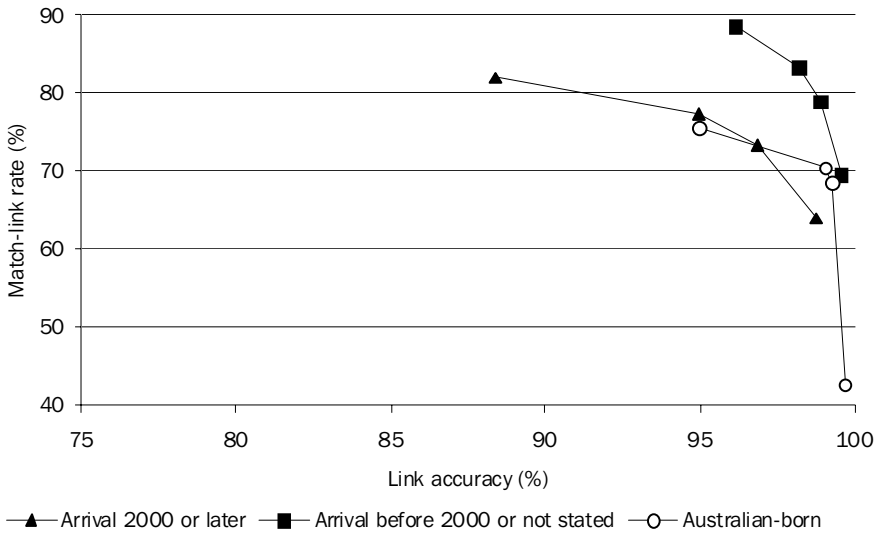
In the future, DIAC may wish to pursue a statistical study which uses the SDB linked to the SLCD when there are at least two waves of Census data. The aim would be to consider research questions like those mentioned in Section 5.5 but with longitudinal data so that longer term outcomes could be investigated.

The feasibility and quality of linking of the SDB to the SLCD will depend not only on the quality of the linking of the SDB to the SLCD sample from the Census, but also on the feasibility and quality of the linking of migrants from Census to Census.

A separate quality study conducted as part of the Census Data Enhancement project assessed the feasibility of forming the SLCD without using name and address as linking variables and the likely quality of the linked data. The formation of the SLCD was simulated by linking the 2006 Census with the 2005 Census Dress Rehearsal conducted one year earlier (the CDR_Census linkage). Details of the linking method can be found in Solon and Bishop (2009).

A brief analysis of the quality of the linking of people born overseas, with year of arrival 2000 or later, in CDR_Census, was conducted to predict the likely quality of the Census to Census component of the potential SDB_SLCD linkage. However, because the Census includes many migrants on temporary visas, it is not possible to tell with much certainty whether a CDR record belonging to a migrant is also on the SDB, so these results should be treated as indicative at best.

6.7 Linkage quality for recent arrivals, earlier migrants and people born in Australia in the CDR_Census linked dataset



Note: For the CDR_Census linkage, four Bronze Standard linkages were created, hence the four points for each category in figure 6.7.

Figure 6.7 shows that the linkage quality in the CDR_Census for migrants who arrived in Australia during or after 2000 was lower than that for other people born overseas (year of arrival before 2000 or not stated). Recent arrivals have higher match-link rates, but lower link accuracy, than persons born in Australia. One reason for these trends is that country of birth is used as a linking variable and this has more distinguishing power when the country of birth is not Australia. The difference between recent and not-so-recent migrants can only be guessed but social attachment to Australia, and English proficiency, are two possibilities that might result in a better completed Census form.

Extensive analysis has been conducted on the CDR_Census dataset to assess quality issues. The full assessment of the likely quality of the SLCD can be found in Bishop (2009). Chipperfield (2009) provides results for one important aspect of this assessment.

It has been recommended that when linking without name and address, a very low cut-off be used. In figure 6.7 this corresponds to a link accuracy of 88.4% and a match-link rate of 82.0% for recent migrants, and a link accuracy of 96.1% and a match-link rate of 88.6% for longer term migrants. These match-link rates compare very favourably with response rates obtained in successive waves of longitudinal surveys. This is discussed further in Section 7.

In summary, the evidence points to an SLCD that will be of sufficient quality to allow linking with the SDB.

7. IMPLICATIONS FOR CONDUCTING A STATISTICAL STUDY

This quality study has assessed the feasibility and likely quality of linking the SDB to the SLCD. It is outside the scope of this quality study to disseminate linked datasets or tabular aggregate data.

The findings from the quality study can be used to assist in making a decision on whether it is worthwhile conducting a statistical study in the future. The statistical study would be created by linking the SDB to the SLCD (5% of Census), not using name and address (Bronze Standard linkage), and the restrictions placed on the dataset would be consistent with all outputs produced by the ABS.

There are two main issues relating to the usefulness of a future statistical study that links the SDB to the SLCD. These are:

- The level at which reliable estimates can be produced because the SLCD is only a 5% sample of the full Census; and
- The quality of the linking and how this affects which analyses are possible or not.

When considering these issues, it is important to keep in mind the other statistical information which is currently available (from other surveys and administrative datasets), and whether or not a statistical study would facilitate new types of analyses or greater accuracy. An assessment can then be made about the usefulness of a statistical study.

In regards to the first issue listed above, Section 6.3 discussed the sampling error introduced by using a 5% sample of the full Census. Although some finer level analyses would become unreliable, it has been demonstrated that the sample size of the SDB_SLCD linked file is larger than any existing similar datasets, and would allow production of estimates at a more detailed level than are available from other sources. That is, the sampling error does not severely restrict the usefulness of the SDB_SLCD linked file in terms of the information gaps it could fill.

Regarding the quality of the linking (second issue listed above), the results presented in this report indicate that linking the SDB to the SLCD is feasible and can produce useful information. However, there are known quality issues that need to be raised when reporting conclusions, and in some instances quality issues may mean it is not appropriate to produce output for specific subpopulations. A thorough understanding of the quality issues raised in this quality study is essential for correctly interpreting and presenting analysis of the linked data.

The major quality issue is the number of SDB records in the population of interest that cannot be linked. This issue was discussed in detail in Sections 5.1 and 5.2. To draw conclusions from the linked data about the entire population of interest assumes that the unlinked records would have similar characteristics. In terms of the characteristics (variables) listed on the SDB (e.g. age, sex, visa category), the linked

and unlinked files generally had similar proportional distributions, but there were some subpopulations that had a small degree of under-representation on the linked file (see Section 5.4 for more details).

The other issue to consider is whether or not the analysis variables from the Census (e.g. employment status) have similar proportional distributions for the linked and unlinked records, although this is hard to assess since the unlinked SDB records do not have the Census variables. It is expected that there would be some correlation between the observed SDB variables and the unobserved Census variables for the unlinked records. Thus the fact that unlinked records have generally similar distributions of SDB characteristics to the linked records, indicates that the Census variables may also be similar.

In terms of SDB records that could be linked, migrants on skilled visas, particularly from Asia, had consistently lower quality measures (match-link rate, link accuracy) compared to migrants on other visa categories. This means that conclusions drawn from analysis of this subpopulation would be potentially more affected by linking error than other migrants.

Regarding the choice of Bronze Standard to use for a statistical study, as discussed in Section 5.6, it is recommended that Bronze Low be the standard of choice. From the analysis presented in this paper, and also from extensive analysis of linked data from another CDE quality study, it has been found that missed matches result in more bias, and thus cause more problems with drawing conclusions from linked data, than do incorrect links.

Considering that the Bronze Low Standard linked 65.6% of the SDB records, this compares favourably with DIAC's *Longitudinal Survey of Immigrants to Australia* (LSIA 3) which for wave 1 of the survey (six months after arrival) had a response rate of roughly 48%. The *Labour Force Status and Other Characteristics of Recent Migrants Survey* and *General Social Survey* had response rates of 95% and 86.5% respectively, but they both have much smaller sample sizes and are not longitudinal surveys.

Considering all the information presented above, the results from this quality study indicate that linking the SDB to the SLCD is feasible and can produce useful information that no other data source currently provides. However, it is essential that any user of the linked data has a thorough understanding of the quality issues identified in this quality study, to ensure that the linked data are correctly interpreted and appropriately used.

These issues could be somewhat mitigated by making improvements in the quality of data being fed into the linking process. The ABS and DIAC will continue to collaborate to identify and implement improvements to the SDB for linking and broader statistical research purposes. In particular, improvements to the completeness and timeliness of SDB address information would enable more links to be made and thus reduce the uncertainty in conclusions associated with unknown characteristics of unlinked records.

8. REFERENCES

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

—— (2006a) *Enhancing the Population Census: Developing a Longitudinal View*, Discussion Paper, cat. no. 2060.0, ABS, Canberra.

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

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

—— (2006d) *General Social Survey: Summary Results, Australia*, cat. no. 4159.0, ABS, Canberra.

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

—— (2007b) *Deaths, Australia, 2007*, cat. no. 3302.0, ABS, Canberra.

—— (2007c) *Labour Force Status and Other Characteristics of Recent Migrants, Australia, November 2007*, cat. no. 6250.0, ABS, Canberra.

Bishop, G. (2009) "Assessing the Likely Quality of the Statistical Longitudinal Census Dataset", *Methodology Research Papers*, cat. no. 1351.0.55.026, 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.

Department of Immigration and Citizenship (2009) *Longitudinal Survey of Immigrants to Australia*, DIAC, Canberra, (last viewed 5 August 2009) <http://www.immi.gov.au/media/research/lsia/>

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.

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