1351.0.55.041

Australian Bureau of Statistics

Research Paper

# Assessing the Quality of Linking School Enrolment Records to 2011 Census Data

New Issue

Research Paper

# Assessing the Quality of Linking School Enrolment Records to 2011 Census Data

National Centre for Education and Training

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) TUES 30 APR 2013

ABS Catalogue no. 1351.0.55.041



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 Andrew Webster, National Centre for Education and Training, on Canberra (02) 6252 6228 or email <education.statistics@abs.gov.au>.

# ASSESSING THE QUALITY OF LINKING SCHOOL ENROLMENT RECORDS TO 2011 CENSUS DATA

National Centre for Education and Training

## EXECUTIVE SUMMARY

As part of the Australian Bureau of Statistics' Census Data Enhancement project, this Education Quality Study was conducted to assess the feasibility of linking government school enrolment records to the 2011 ABS Census of Population and Housing, with and without the use of name and address as linking variables. The primary aim of the study was to establish the quality of linked datasets where name and address are not used for linkage[1] – Bronze linkage.

Government school enrolment data (non-financial information) at the unit record level were provided to the ABS by the education departments from Queensland, South Australia, Tasmania and the Northern Territory. The datasets supplied for the CDE Education Quality Study covered the years 2010 and 2011.

Preparation of the datasets for linkage involved standardisation (making the two datasets more compatible) and deduplication (removing multiple records). In particular, substantial effort was made to ensure that students had valid address data, which could be used to generate statistical geography codes (Mesh Block or Statistical Area 1).

Probabilistic[2] and deterministic[3] methods of linkage were then used to integrate the datasets. This report examines the probabilistic methods. A research paper discussing the outcomes from the deterministic linkage is planned for release later this year (2013).

For each participating jurisdiction, the integrated dataset based on name and address (together with other linkage variables) – the Gold dataset – was generated using probabilistic linkage and clerical review.[4] Over 85% of school enrolment records were integrated with an equivalent Census record in the Gold linkage for Queensland, South Australia and Tasmania. The Northern Territory Gold linkage rate was lower at 76%.

1 Address information was used to generate statistical geography levels (Mesh Block, Statistical Area 1 and 2). Statistical geography (not exact address) was used for the Bronze linkage.

2 Probabilistic linking compares records from two datasets using several variables common to both datasets and generates a single numerical measure of how well two particular records match. This allows ranking of all possible record pairs and assignment of the optimal link. For more information, see Section 4.2.

3 Deterministic linking compares only record pairs that match exactly or almost exactly (e.g. age within one year) on a combination of variables, seeking unique matches wherever possible.

4 Clerical review is the manual examination of record pairs by a person to determine whether the pairs are links or non-links.

The Bronze probabilistic linkage used two standards – the first with a high level of accuracy (Bronze high) and the second aiming to maximise the linkage rate (Bronze low). While the linkage accuracy of the Bronze low standard is lower than for the Bronze high, all the extra links assigned match on age, sex and small geographic area (Mesh Block or Statistical Area 1).

The quality of the linked datasets was assessed on the basis of four indicators:

1. expected links – the number of links expected to be made after taking the net undercount in the Census into account;
2. unlinked records – a short analysis of the types of missing data on the unlinked records;
3. linkage accuracy and match-link rate – the degree to which the Bronze datasets approach the accuracy and coverage of the Gold datasets; and
4. representation of population sub-groups on the linked datasets – a series of tables comparing relative frequencies from the school enrolment records with the Gold and Bronze linked datasets on the basis of various socio-demographic and geographic characteristics.

The results from the CDE Education Quality study indicate that linking government school enrolment records, and education records more generally, to the Census using probabilistic linkage methods without name and address is feasible. The study has produced linked datasets that, in general, link a high proportion of school enrolment records to the Census and accurately link equivalent records (that is the same individual from each dataset). While this is particularly the case for Gold linkage, the Bronze datasets also display these qualities of coverage and accuracy. As a result, the linked datasets are highly representative of the enrolled school population. The majority of students in the enrolment records were present in the linked datasets, and were well represented across a range of demographic, geographic and social characteristics. The linked datasets could support a range of investigations, including analysis of pathways taken by school leavers in the year after school, transitions from government to non-government schooling, differential Indigenous status across the two collections, and the socio-economic characteristics of students and parents.

Additional research papers are planned which will further examine the quality of the linked datasets and their analytical potential.

# CONTENTS

# LIST OF TABLES AND FIGURES

# ACKNOWLEDGEMENTS

The 2011 Census Education Quality Study was funded by the Strategic Cross-sectoral Data Committee. This committee, which operates within the governance structure under the Ministerial Standing Councils in the fields of education and training, has responsibility for a work program supporting the integration of cross-sectoral education data, and improving longitudinal and outcomes information.

The ABS acknowledges assistance provided through Memoranda of Agreement with the Queensland Department of Education, Training and Employment; the South Australian Department for Education and Child Development; the Tasmanian Department of Education and the Northern Territory Department of Education and Training.

This paper was prepared in the ABS National Centre for Education and Training Statistics, with assistance from the ABS Analytical Services Unit. The principal authors and researchers were Noel Hansen, Elaine Lay, Presley Peter, Luxshme Ranjan, Rita Scholl, Felicity Splatt, Caitlin Szigetvari and Andrew Webster.

This project has benefitted from the expert and diligent contributions made by officers from many teams in the ABS. The authors would like to acknowledge and thank the ABS Analytical Services Unit, in particular Sean Buttsworth, Paul Campbell, Phillip Gould, Damien Melksham, Peter Rossiter and Gokay Saher for their guidance throughout the study; Jessica Newton from ABS Geography; Merryn Barnard, Brent Bufton, Gennevene Ensor, Alina Harabor, Mary Jackson, and Libby O'Toole of the ABS National Centre for Aboriginal and Torres Strait Islander Statistics; and the ABS Data Linkage Centre, particularly Divya Dass, Brendan Kelly, Neetu Mittal and Bradley White.

# ASSESSING THE QUALITY OF LINKING SCHOOL ENROLMENT RECORDS TO 2011 CENSUS DATA

National Centre for Education and Training

## ABSTRACT

As part of the Australian Bureau of Statistics' Census Data Enhancement project, this Education Quality Study was conducted to assess the feasibility of linking government school enrolment records to the 2011 ABS Census of Population and Housing, with and without the use of name and address[1] as linking variables. This initial paper details the methodology used in the linkage process, the outcomes of the project and the quality of the resultant datasets.

The results from this quality study indicate that linking government school enrolment records, and education records more generally, to the Census using probabilistic linkage methods without name and address is feasible. This work has produced linked datasets which could be used for a range of investigations, including pathways taken by school leavers in the year after school, transitions from government to non-government schooling, differential Indigenous status across the two collections, and the socio-economic characteristics of students and parents. Additional research papers are planned which will further examine the quality of the linked datasets.

1 Address information was used to generate statistical geography levels (Mesh Block, Statistical Area 1 and 2). Statistical geography (not exact address) was used for the Bronze linkage.

# 1. INTRODUCTION

Commencing with the 2006 Census of Population and Housing, the ABS began investigating ways to enhance the value of Census data by combining it with other datasets using statistical data integration techniques. The purpose of integrating the Census with other data sources is to gain more information from the combination of datasets than is available from the datasets separately; without increasing the burden on providers through further survey collections. The ABS has embarked on an extended range of Census Data Enhancement (CDE) projects using data from the 2011 Census. These projects will expand the range of official statistics available to Australian society, and improve the evidence base used to underpin good government policy, program management and service delivery.

The 2011 CDE suite of projects includes statistical data integration of:

1. 2011 Census data with a small number of predetermined datasets during the Census processing period using name and address, to investigate linkage quality
2. 2011 Census data with a small number of predetermined datasets during the Census processing period using name and address, to create statistical outputs
3. 2006 Census data with 2011 Census data without name and address to create Wave 2 of the 5% Statistical Longitudinal Census Dataset (SLCD)
4. the SLCD with other datasets without using name and address for statistical and research purposes
5. 2011 Census data with other datasets without using name and address after the Census processing period.

The 2011 CDE quality studies comprised the Education Quality Study and the Migrants Quality Study. The Education Quality Study aimed to test the quality of linked datasets using the 2011 Census of Population and Housing combined with data from 2010 and 2011 government school enrolment collections from four participating jurisdictions: Queensland, South Australia, Tasmania and the Northern Territory.

The Migrants Quality Study replicates a corresponding study that was conducted after the 2006 Census. It aims to test the quality of linking the 2011 Census with the Settlements Database held by the Commonwealth Department of Immigration and Citizenship (see ABS, 2010). It is anticipated that the linked dataset will contain reliable and consistent information with respect to the labour force status, education, income, housing and caring arrangements of recently arrived migrants and will assist in assessing and improving migration policies and programs. Results from the Migrants Quality Study will be released later this year (2013).

Features of these studies include:

1. linked datasets are created during the Census processing period when name and address are available
2. the linked datasets created through these projects do not leave the ABS and are only accessible to those ABS officers directly involved in the study
3. once the purpose of each study has been met, all linked datasets that were integrated using name and address are destroyed.

The primary aim of the 2011 Education Quality Study was to establish the quality of linked Census and education datasets where name and address data are not used for linkage[2]. This is achieved by comparing the results of a dataset which is linked using name and address (the Gold dataset) with corresponding datasets which are linked without using name and address (Bronze datasets). Understanding the quality of the linkage is important because it will shape the linkage requirements and best practice for future integration projects involving education enrolment information. A long-term goal in the education and training sector is to develop a data integration strategy that will increase our understanding of the pathways and outcomes of Australian students from early childhood education to schooling, post-school education and ultimately to the labour market. Longitudinal analysis of this information could support a better understanding of the underlying factors affecting student progress and outcomes, including workforce participation.

2 Address information was used to generate statistical geography levels (Mesh Block, Statistical Area 1 and 2). Statistical geography (not exact address) was used for the Bronze linkage.

## 2. THE 2011 CDE EDUCATION QUALITY STUDY

There is considerable demand from government agencies, researchers and the wider community for data that assists in understanding educational pathways and outcomes for students. The availability of relevant and high-quality information is crucial for decision making in education.

Globally, data linkage is seen as a cost effective and comprehensive way to meet this demand, without increasing respondent burden. Researchers in Canada, for instance, have linked data from the Early Development Index (students are tested for a range of vulnerabilities in Kindergarten) and standardised literacy and numeracy testing in the fourth year of schooling. This linked dataset has enabled the identification of regions where students have been unable to remedy a negative start in school or need support to maintain success through their early school years. This subsequently allows government to target funding and interventions in regions where it is most needed to help students towards positive schooling outcomes (Lloyd and Hertzman, 2009). The need for similar linkage activity in Australia using the corresponding collections, the Australian Early Development Index (AEDI) and the results of the NAPLAN testing program has been flagged by a variety of stakeholders in discussions with the ABS.

There are many sources of national statistics about school students in Australia for which data integration offers the potential for enriched multidimensional information. These include: the National Schools Statistics Collection (NSSC), the Longitudinal Surveys of Australian Youth (LSAY), the Longitudinal Study of Australian Children (LSAC) and international collections such as the Program for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS). In addition, the Census of Population and Housing, which collects information on all persons in Australia on a five-yearly cycle, includes an item on school sector (government, Catholic, Independent) and a wide range of socio-demographic information. For those students still living with parents or guardians, characteristics of the family can be derived (e.g. parental education, parental occupation, family income).

Data linkage, while much more cost effective than a survey, requires extensive preparation, in the repair and standardisation of the raw data, and considerable statistical and technical nous in the execution of the linkage itself (Solon and Bishop, 2009). Where the original datasets comprise data for a sub-population only, such as students, the range of possible variables for linkage (and the range of values for those variables), is constrained and it becomes more difficult to match two equivalent records because the applicable records are much more similar; and less distinct (see Bishop, 2009: Section 3.3).

In the CDE Education Quality Study, the ABS, with the assistance of data providers from four participating jurisdictions, has taken on the challenge of constructing linked datasets that address needs for data on the socio-economic context of school students. The study links government school enrolment records from Queensland, South Australia, Tasmania and the Northern Territory with the equivalent Census records from students and their parents or caregivers, with the aim of creating a rich data source for the analysis of the personal and family characteristics of school students, and outcomes for students in the first year after leaving school. Potential future linkages between the Census, early childhood data (AEDI) and achievement data (NAPLAN) could further reveal how family characteristics impact on school readiness, achievement and long-term outcomes for students.

The 2011 Census processing period provided the opportunity to link the full Census with school enrolment records using two approaches, one with and the other without name and address as linking variables. Linking with name and address (Gold linkage) provides a benchmark for assessing linkage quality when name and address are not available (Bronze linkage). It should be noted that all names and addresses collected as part of the Census were destroyed once Census processing was completed, so that all potential future linkage between the 2011 Census and other datasets will be performed using Bronze linkage techniques.

# 3. THE DATA

This section provides an overview of the two data sources being brought together for linkage; that is government school enrolment data and the 2011 Census. The data quality issues which impact on the compatibility of the two sources for linkage purposes are then discussed.

## 3.1 Government school enrolment data

For this study, government school enrolment data (non-financial information) at the unit record level were provided to the ABS by the education departments from Queensland, South Australia, Tasmania and the Northern Territory. The datasets covered the years 2010 and 2011.

The data differed slightly for each jurisdiction, but contained common variables about the student, the school and the students' parents or caregivers. Within some jurisdictions, the datasets also changed slightly between 2010 and 2011 in terms of how data were collected and/or coded. Table 3.1 summarises the number of enrolment records received from each jurisdiction.

**3.1 Number of enrolment records, by jurisdiction and year (2010–2011)**

| | *Queensland* | *South Australia* | *Tasmania* | *Northern Territory* |
|---|---|---|---|---|
| Year | | | | |
| 2010 | 492,656 | 170,804 | 61,996 | 29,689 |
| 2011 | 496,778 | 169,420 | 61,497 | 29,804 |
| Total records | 989,434 | 340,224 | 123,493 | 59,493 |

Source: Government school enrolment records, 2010–2011.

## 3.2 Census data

The 2011 Census dataset used for this study consisted of 20,928,304 records, excluding imputed records. Imputed records are created to account for people for whom no Census form was returned – see the *Census Dictionary* (ABS, 2011b) for more information about imputation. Almost two million people (1,955,826) were identified as attending a government school. Almost one third of these, 624,151 people (32%), were from the states and territory participating in the study: Queensland, South Australia, Tasmania or the Northern Territory (usual residence). This study used the full Census dataset for linkage, enabling potential linkages for students who had moved interstate or left the government schooling system during the period of the study. Table 3.2 shows the Census data on the number of students attending government schools by jurisdiction.

**3.2 Number of students attending government educational institutions, by jurisdiction and school type, 2011**

| | *Queensland* | *South Australia* | *Tasmania* | *Northern Territory* |
|---|---|---|---|---|
| Attending a government school | | | | |
| Infants / Primary | 258,616 | 87,778 | 30,054 | 15,171 |
| Secondary | 154,133 | 52,221 | 18,175 | 8,003 |
| Total | 412,749 | 139,999 | 48,229 | 23,174 |
| Type of educational institution not stated | 311,791 | 98,694 | 31,624 | 26,242 |

Cells in this table have been randomly adjusted to avoid the release of confidential data.
Source: ABS Census of Population and Housing, 2011.

## 3.3 Preparing data for linkage

Preparing data for linkage involves several steps taken to increase the compatibility of the original datasets and to identify and address any data quality issues. These steps included:

- ensuring that the population of interest is included in both datasets
- ensuring the time frame of collection of both datasets is comparable
- identifying variables suitable for linkage and ensuring these variables are collected and coded in a compatible way or recoding variables where necessary to maximise the possibility for linkage to occur
- investigating missing, invalid or conflicting data and its potential impact on linkage and repairing this data where possible
- ensuring there is only one record per person on each dataset.

Each step in the preparation process is explained in this section.

### *3.3.1 Scope of the student population*

Not all government school enrolment records are generated by a student entering a public school to attend a standard grade from Kindergarten / Preparatory year to Year 12. An enrolment record is also generated when a child is enrolled in a pre-school program run by a government school and by students who are completing Vocational Education and Training (VET) programs delivered at a public school. These students were not flagged or identified, so they are included in the linked data. Student age gives some indication as to the small proportion of records that are enrolled in a government school, but not necessarily studying school curricula. Table 3.3 shows the proportion of students whose age is outside the expected range for school students for each state.

There are also a very small proportion of school students who would not have an equivalent Census record available for linkage. These are discussed in more detail in Section 5.1.

**3.3 Age distribution of students, by jurisdiction**

| | *Queensland* | *South Australia* | *Tasmania* | *Northern Territory* |
|---|---|---|---|---|
| Age (%) | | | | |
| 0–3 years | 0.2 | – | – | – |
| 4–20 years | 99.7 | 96.5 | 97.3 | 99.6 |
| 21 years or over | 0.2 | 3.5 | 2.7 | 0.4 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 |

– nil or rounded to zero (including null cells)
Source: Government school enrolment records, 2010–2011.

### *3.3.2 Time frame comparability*

The ability to link records from two datasets is maximised when the data is collected in the same time frame. This is rarely possible. Most data linkage occurs between sources that are collected in different time frames, and the greater this difference, the less likelihood of a successful linkage. The two types of dataset linked in the CDE Education Quality Study have different time frames.

The school enrolment data is an administrative dataset. Usually, parents / caregivers only complete an entire school enrolment form when a student starts or changes school. As a result, the data from the school enrolment records can be several years old. Parents / caregivers are asked annually to update a subset of information (their contact details), and the student's grade level and other school characteristics are also regularly updated. Since the introduction of annual performance reporting, however, which releases student outcomes information by characteristics collected on the school enrolment form (especially parental background information), there has been substantial improvement in the timeliness of information held by school administrative systems.

While school enrolment data is collected over decades and updated incrementally, Census data is a 'snapshot' – with the vast majority of Census forms being completed on Census night (8 August, 2011).

As an example of the currency of school enrolment information, the Queensland dataset included the date on which the enrolment form was completed. For almost 14% of the Queensland 2011 cohort, the school enrolment form was collected more than five years ago. Table 3.4 shows the age of the enrolment records for the Queensland 2011 cohort.

**3.4 Year of completion of enrolment form, Queensland, 2011 cohort**

| | *No.* | *%* |
|---|---|---|
| Year of completion | | |
| 2011 | 133,534 | 27.0 |
| 2010 | 106,019 | 21.4 |
| 2009 | 81,483 | 16.5 |
| 2008 | 62,086 | 12.6 |
| 2007 | 44,309 | 9.0 |
| 2006 | 27,331 | 5.5 |
| 2005 | 20,433 | 4.1 |
| 2004 | 14,509 | 2.9 |
| 2003 | 2,303 | 0.5 |
| 2002 | 1,167 | 0.2 |
| 2001 or earlier | 1,263 | 0.3 |
| Total | 494,437 | 100.0 |

Source: Queensland government school enrolment records, 2011.

Address one year ago and five years ago are collected in the Census, providing a potential remedy for students whose enrolment records may not have fully reflected the student's details at the time of the Census but had been completed or updated within the last five years. Mismatched time frames are more damaging to linkage where fewer linkage variables are available, particularly if name and address is not available (Bronze standard linkage).

### *3.3.3 Selecting and standardising variables for linkage*

The school enrolment and Census records have many variables in common. In addition to personal details, such as name and address, both datasets use the ABS standard Indigenous status question, and Australian standards for classifying country of birth and main language spoken at home. As a result of the data standards work implemented in 2006–2008 by the (then) Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA), school enrolment forms across Australia have standard ways of collecting parental information, including sex, Indigenous status, country of birth, main language spoken at home, occupation type and educational attainment (see ACARA, 2012). Apart from occupation type (which needs to be aligned with the Census occupation classification), these items are all collected in a compatible way on school enrolment forms and the Census. As a result, there were many common variables available for use as linkage variables (see table 4.2).

While many of these variables were directly comparable, some variables required adjustment to optimise the likelihood of linkage. For instance, as would be expected, the country of birth recorded for the majority of students was Australia and their main language spoken at home was English. The lack of variability for these variables decreased their value as linkage variables. As a result, Australia and English were

changed to 'missing' to prevent these variables from biasing probabilities used for the linkage. Also, Indigenous status was not used for linkage, to avoid bias in subsequent analysis of Indigenous identification across the school enrolment and Census datasets.

Substantial effort was made to ensure that name fields were standardised across datasets, for instance accurately identifying the first name when more than one given name is provided. The address variables on the school enrolment records also required extensive processing in order to identify a valid Mesh Block, Statistical Area 1 and Statistical Area 2 for as many students as possible.

### *3.3.4 Multiple records*

Multiple records for a unique person pose an issue when linking across datasets. Duplicate records make it more difficult to extract a match between two equivalent records on the original datasets, as the records are less distinct. Where a match does occur, these records are linked and removed from the pool of possible links. The remaining multiple records continue in the pool of possible links and inevitably become a missed link or a false link. Either alternative is detrimental to linkage quality. If there are unnecessary missed links, the overall linkage rate will be falsely deflated, and if there are unnecessary false links, the link accuracy will be decreased. See Section 5.3 for more information on link accuracy.

As the school enrolment records covered 2010 and 2011, the majority of students had two school enrolment records (referred to as the continuous student population). Multiple enrolment records for a student also arise for several other reasons:

- students may change schools within jurisdictions
- students may attend more than one campus for various educational or family reasons
- system or clerical error.

In addition to the majority of students who were continuously enrolled from 2010 to 2011, there were also students who were only enrolled in 2010 (referred to as the leaving student population), and students who were only enrolled in 2011 (the new student population).

All school enrolment records pertaining to a particular student were identified and concatenated to produce a student-level, rather than record-level, dataset. Unique students were identified using their state or territory student identification number and also by probabilistically linking each dataset to itself to identify students with more than one student identification number (proportionally this was a very small number of students in each jurisdiction). Table 3.5 summarises the number of records in the student-level dataset in each jurisdiction.

**3.5 Number of unique enrolment records, by jurisdiction and year of enrolment, 2010–2011**

| | *Queensland* | *South Australia* | *Tasmania* | *Northern Territory* |
|---|---|---|---|---|
| Enrolled in both 2010 and 2011 ('continuous students') | 419,483 | 143,428 | 50,595 | 22,641 |
| Enrolled in 2010 only ('leaving students') | 71,620 | 26,262 | 8,873 | 6,260 |
| Enrolled in 2011 only ('new students') | 74,954 | 25,076 | 8,925 | 6,371 |
| Unique students | 566,057 | 194,766 | 68,393 | 35,272 |

Source: Government school enrolment records, 2010–2011.

After creating the student-level school enrolment files, the number of unique records was checked for accuracy and completeness against the Census and the National Schools Statistics Collection (NSSC). Table 3.6 shows the variation in total numbers between the study data and the other two sources. The availability of name and address data for this study helped to identify more multiple records than is possible for the NSSC without this identifying information. Census counts for government schools are lower than for the NSSC or the enrolment datasets used for this study because the Census data for educational participation generally contains a high rate of not stated for Institution Type (22–36% for these jurisdictions).

**3.6 Government school participation, by source and jurisdiction, 2011**

| | *Queensland* | *South Australia* | *Tasmania* | *Northern Territory* |
|---|---|---|---|---|
| CDE Education Quality Study | 494,437 | 168,504 | 59,520 | 29,012 |
| Census counts | 412,749 | 139,999 | 48,229 | 23,174 |
| National Schools Statistics Collection | 496,275 | 168,104 | 59,536 | 29,343 |

Sources: Government school enrolment records, 2011; ABS Census of Population and Housing, 2011; National School Statistical Collection, 2011.

### *3.3.4 Missing, invalid and conflicting data*

Missing, invalid or conflicting values were identified during the interrogation of the original datasets and the construction of the student level records for a small proportion of students. To remedy this, some values (sex and date of birth) were checked and repaired by the data providers.

If other values were missing, invalid or conflicting between multiple records, a single value was selected for affected students for the purpose of linkage. Where the student had a 'home' (main) school, then the value from that record was selected. However, if the student had multiple, conflicting records and several or no home schools, then the most recent value was selected for the student level datasets.

This process included the resolution of conflicting values for Indigenous status. As a result, it is possible that a student who identified as Aboriginal and/or Torres Strait Islander at some stage of their schooling may be treated as non-Indigenous if they identified as non-Indigenous on their most recent enrolment record. The reverse is also possible; some students may be identified as Aboriginal and/or Torres Strait Islander who had previously identified as non-Indigenous.

Finally, where a student had multiple values for Country of birth and Main language spoken at home, the most common value other than Australia and English respectively was given preference in selecting values for linkage.

Table 3.7 shows the proportion of missing and invalid data for the variables used to link the school enrolment records with the Census, after standardisation (see Section 4.1).

**3.7 Missing data, for Gold linkage variables, by jurisdiction, school enrolment records**

| | *Jurisdiction* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Queensland* | | *South Australia* | | *Tasmania* | | *Northern Territory* | |
| *Linking variable* | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| Name information | | | | | | | | |
| First name | 38 | – | 11 | – | – | – | 4 | – |
| Surname | 123 | – | 28 | – | 11 | – | 26 | 0.1 |
| Age-related information | | | | | | | | |
| Day of birth | – | – | – | – | – | – | – | – |
| Month of birth | – | – | – | – | – | – | – | – |
| Year of birth | – | – | – | – | – | – | – | – |
| Personal characteristics | | | | | | | | |
| Sex | – | – | – | – | – | – | – | – |
| Ethnicity | | | | | | | | |
| Country of birth | 59,209 | 10.5 | – | – | 5,598 | 8.2 | 2,167 | 6.1 |
| Main language spoken at home(a) | 1,285 | 0.2 | 170,409 | 87.5 | 5,097 | 7.5 | 7,007 | 19.9 |
| Address Information | | | | | | | | |
| Statistical Area 1 | 15,096 | 2.7 | 18,757 | 9.6 | 1,917 | 2.8 | – | – |
| Mesh Block | 28,714 | 5.1 | 21,989 | 11.3 | 2,321 | 3.4 | 4397 | 12.5 |
| Street number | 79,541 | 14.1 | 23,859 | 12.3 | 1,796 | 2.6 | 10,778 | 30.6 |
| Street name | 13,050 | 2.3 | 16,841 | 8.6 | 1,377 | 2.0 | 9,566 | 27.1 |
| Suburb | 204 | – | 11,118 | 5.7 | 14 | – | 360 | 1.0 |
| Postcode | 196 | – | 11,106 | 5.7 | 137 | 0.2 | 349 | 1.0 |
| *Total* | *566,057* | *100.0* | *194,766* | *100.0* | *68,393* | *100.0* | *35,272* | *100.00* |

– nil or rounded to zero (including null cells)
(a) In South Australia, the school enrolment form is intended to collect Main language other than English. As a result, Main language spoken at home is missing for most students.

For the Census data, missing data in the date of birth variables was particularly detrimental to linkage, especially for the Bronze linkage, where name and address could not mitigate if this data was missing.

Table 3.8 shows the missing data for the linkage variables for people aged 5–15 years in Queensland, South Australia, Tasmania and the Northern Territory. (Name and address is not provided in this table as it is no longer available.)

### 3.8 Missing data, for Gold linkage variables, by jurisdiction, Census records, 5–15 year olds

| | *Jurisdiction* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Queensland* | | *South Australia* | | *Tasmania* | | *Northern Territory* | |
| *Linking variable* | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| Age-related information | | | | | | | | |
| Day of birth | 80,842 | 13.1 | 23,502 | 11.4 | 8,221 | 12.2 | 7,130 | 22.1 |
| Month of birth | 80,775 | 13.1 | 23,475 | 11.4 | 8,219 | 12.2 | 7,127 | 22.1 |
| Year of birth | – | – | – | – | – | – | – | – |
| Personal characteristics | | | | | | | | |
| Sex(a) | 5,562 | 0.9 | 1,927 | 0.9 | 876 | 1.3 | 355 | 1.1 |
| Ethnicity | | | | | | | | |
| Country of birth | 10,772 | 1.7 | 3,352 | 1.6 | 776 | 1.1 | 651 | 2.0 |
| Main language spoken at home(a) | 6,347 | 1.2 | 2,224 | 1.1 | 614 | 0.9 | 892 | 2.8 |
| Address Information | | | | | | | | |
| Statistical Area 1 | 4,393 | 0.7 | 1,171 | 0.6 | 158 | 0.2 | 227 | 0.7 |
| Mesh Block | 4,393 | 0.7 | 1,171 | 0.6 | 158 | 0.2 | 227 | 0.7 |
| *Total Census records* | *615,767* | *100.0* | *206,273* | *100.0* | *67,507* | *100.0* | *32,263* | *100.00* |

– nil or rounded to zero (including null cells)
(a) Prior to imputation.

# 4. THE LINKAGE PROCESS

This section provides an overview of the work undertaken by the ABS to create school enrolment to Census linked datasets for Queensland, South Australia, Tasmania and the Northern Territory.

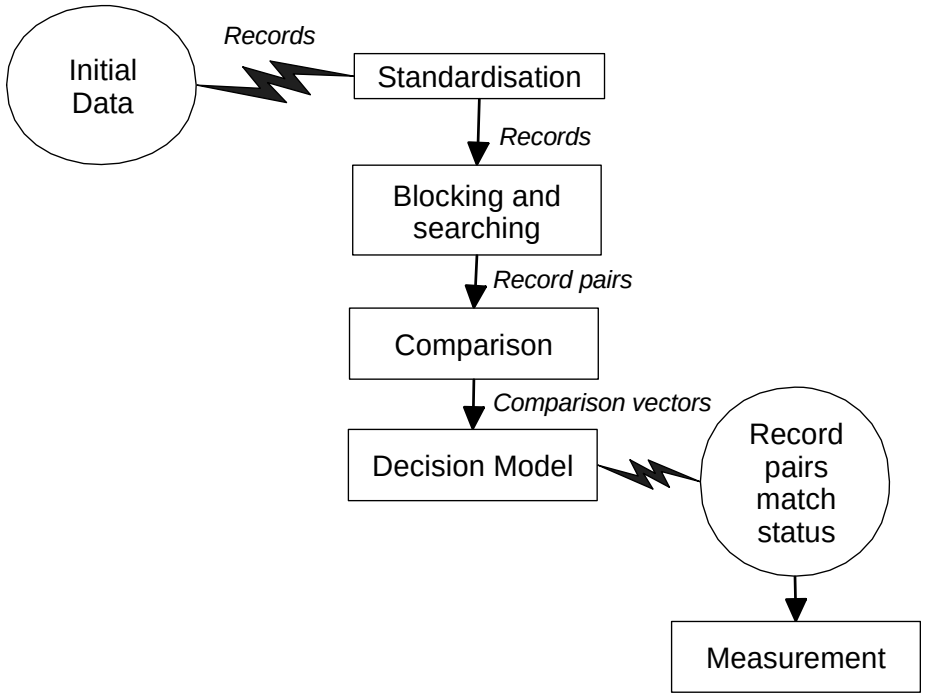
## 4.1 Linking methodology

The linking methodology examined in this paper to link the school enrolment records and the 2011 Census, either with or without name and address, is called probabilistic linking. This method links records from two datasets using several variables common to both datasets. 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 (Solon and Bishop, 2009).

The probabilistic linking methodology used here can be generalised into the following steps:

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

4.1 The generalised data linkage process



***Standardisation***. Before records on the two datasets are compared, the contents of the two datasets need to be standardised to facilitate comparison. This includes a number of steps such as verification, recoding and reformatting fields, and parsing text fields. Additionally, some fields require substantial repair. For instance, a first name field may undergo a number of operations, such as the removal or recoding of non-alphabetic characters (hyphens and some blank spaces excepted), search and removal of common prefixes (Mr, Ms) and suffixes (Jr). Names may also undergo nickname standardisation or indexing to a name dictionary.

Some variables differ between the two datasets in a predictable way, and an adjustment is required to negate this difference. For instance, if the two input datasets were collected one year apart, age should be adjusted by one year on one of the datasets to ensure it matches with the other. Some variables are coded differently at different points in time, and concordances may be necessary to create variables which align on the two datasets.

Variables may also be recoded or aggregated in order to obtain a more robust form of the variable. For example, Country of Birth is coded at a fine level on Census data, but may not be reported consistently at this fine level on either the Census or the enrolments records. This means a new variable capturing Country of Birth at a higher level may be required for use in data linking (2 digit instead of 4 digit).

This set of procedures is collectively termed "standardisation". Standardisation takes place in conjunction with a broader evaluation of the dataset, in which potential linking variables are identified. See Section 3.3 for more information about how the datasets were standardised in preparation for linkage.

***Blocking.*** Once data files have been standardised, 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 even moderately large, comparing every record on File A with every record on File B is computationally infeasible. Blocking reduces the number of comparisons by only comparing record pairs where matches are likely to be found – namely, records which agree on a set of blocking variables. Blocking variables are selected based on their reliability and discriminatory power. For instance, sex is partially useful as it is typically well-reported, however it is minimally informative as it only divides datasets into two blocks, and is thus used in conjunction with other variables.

The process of blocking reduces the computational intensity of data linking. However, comparing only records that agree on a particular set of blocking variables means a record will not be compared with its match if it contains missing, invalid or legitimately different information on a blocking variable. To mitigate this, the linking process is repeated a number of times, using a range of different blocking strategies.

For example, on the first pass, a block by a low level of geography (SA1) was used to capture the majority of students who had matching addresses across their enrolments and Census records. This means, however, that those students who had moved since recording their address on their enrolment form were not compared. Records which failed to link in the first pass proceeded to the next pass, in which a different set of blocking variables was used. For the second pass, by blocking on date of birth rather than geography, the students who had moved or who had missing or invalid address information were able to be compared.

***Record pair comparison.*** Within a blocking pass, records on the two files which agree on the specified blocking variables are compared on a number of linking fields. Each linking field has associated *field weights,* which are calculated prior to comparison. Field weights indicate the amount of information (agreement, disagreement, or missing values) a linking field provides about whether the records belong to the same or a different person (true match status). Field weights are based on two probabilities associated with each linking field: first, the probability that the field values agree on a record pair given that the two records belong to the same person (match); and second, the probability that the field values agree on a record pair given the two records belong to different persons (unmatch). These are called $m$ and $u$ probabilities (or match and unmatch probabilities) and are defined as:

$$m = \mathrm{P}\left(\text{fields agree} \mid \text{records belong to the same entity}\right).$$
$$u = \mathrm{P}\left(\text{fields agree} \mid \text{records belong to different entities}\right).$$

Given that the $m$ and $u$ probabilities require knowledge of the true match status of record pairs, they cannot be known exactly, but rather must be estimated. The ABS uses a number of techniques to estimate $m$ and $u$ probabilities. For the series of 2011 linking projects, the Expectation Maximisation (EM) algorithm was used (see Samuels, 2012). In some instances the EM algorithm is deemed unsuitable, or fails to converge on an estimate, and in such cases $m$ and $u$ probabilities are based on those of similar linking projects. As a new feature to the suite of 2011 Census Data Enhancement projects, $m$ and $u$ probabilities for missing data on a linking field were calculated. These capture the probability that a pair belonging to the same individual (match), and a pair belonging to two different individuals (unmatch), are missing on either dataset (or both datasets) for a linking field. Note that $m$ and $u$ probabilities are calculated for each pass, conditional on agreement on the specified blocking fields, as all records compared will agree on blocking variables.

Match ($m$) and unmatch ($u$) probabilities are then converted to agreement, disagreement and missing field weights. The formulae to convert $m$ and $u$ probabilities to field weights are a small extension of the Fellegi and Sunter (1969) linking methodology which now provide weights for missing data.

They are as follows:

$$\text{Agree} = \log_2\left(\frac{m}{u}\right) \tag{4.1}$$

$$\text{Missing} = \log_2\left(\frac{m_{\text{missing}}}{u_{\text{missing}}}\right) \tag{4.2}$$

$$\text{Disagree} = \log_2\left(\frac{1-m-m_{\text{missing}}}{1-u-u_{\text{missing}}}\right) \tag{4.3}$$

Equations 4.1 to 4.3 give rise to a number of intuitive properties of the Fellegi–Sunter framework. First, in practice agreement weights are always positive and disagreement weights are always negative. Second, the magnitude of the agreement weight is driven primarily by the likelihood of chance agreement. That is, a low probability of two random people agreeing on a field (for example, Date of Birth) will result in a large agreement weight applied when two records do agree. The magnitude of the disagreement weight is driven by the stability and reliability of a variable. That is, if a variable is well-reported and stable over time (for example, Sex) then disagreement on the variable will yield a large negative weight. For each record pair comparison, the field weights from each linking field are summed to form an overall record pair comparison weight.

Before calculating $m$ and $u$ probabilities for some variables it is first necessary to define what constitutes agreement. Typical comparison functions include:

- exact match (e.g. sex). Agreement occurs only when the two field values are identical. This criterion is used for most linking fields
- approximate string comparison (e.g. name). Two strings may be said to agree in spite of a certain proportion of missing, differing, or transposed characters, allowing for misspellings, transcriptions of poor handwriting, etc. String comparators (such as Jaro, Winkler comparator) can be used to ensure that both identical and similar string pairs are defined to agree
- numeric difference (e.g. year of arrival). A pair may be defined to agree if their field values differ by an amount less than or equal to a specified maximum difference.

Alternatively, near or partial agreement may be factored into the linking process not in defining agreement but in converting $m$ and $u$ probabilities to weights. For example, the student's age on equivalent records will usually be an exact match, and the $m$ and $u$ probabilities are calculated based on this definition. During linkage, however, a partial agreement weight is given for ages within one year difference.

Blocking fields, linking fields, agreement comparison functions, and $m$ and $u$ probabilities are passed into linking software. Records which agree on the blocking variable(s) are compared on all linking fields.

***Decision model.*** Finally, a decision rule determines whether the record pair is linked, not linked or considered further as a possible link. The first phase of this process is automated, in which a record is assigned to its best possible pairing. This process is known as one-to-one assignment. Ideally (and often true in practice) each record has a single, obvious best pairing, which is its true match. Linking projects in the ABS have typically used an auction algorithm to assign optimally one record on the first dataset to one record on the second dataset. The auction algorithm maximises the sum of all the record pair comparison weights through alternative assignment choices, such that if a record A1 on File A links well to records B1 and B2 on File B, but record A2 links well to B2 only, the auction algorithm will assign A1 to B1 and A2 to B2, to maximise the overall comparison weights for all record pairs.

The second phase of the decision rule stage takes the output of one-to-one assignment and decides which pairs should be retained as links, and which should be rejected as non-links. This is done by defining cut-off weights against which record pair comparison weights are evaluated. The simplest decision rule uses a single cut-off such that all record pairs with a weight greater than or equal to the cut-off 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 employs lower and upper cut-offs. Record pairs with a weight above the upper cut-off are declared links while those with a weight below the lower cut-off are declared non-links. The record pairs with weights between the upper and lower cut-offs are designated for clerical review.

In clerical review, each record pair is assessed by inspection to resolve match status. A clerical reviewer is often able to utilise information which cannot be captured in the automated comparison process, such as variations in names and common transcription errors (e.g. 1 and 7). Reviewed records are either accepted as links or rejected.

Note that even where the original data is of very high quality, equivalent records may not be identical across all the blocking and linkage variables. For this reason, several 'passes' are used to optimise the opportunity for equivalent records to be linked, with different combinations of blocking and linking variables for each pass. Records on each dataset not linked on one pass are included in the pool of possible links for the next pass.

## 4.2 Implementation in the 2011 CDE Education Quality Study

Both probabilistic and deterministic linking methods[3] have been tested as part of this quality study, however, only the probabilistic methods have been assessed in this paper. Two types of probabilistic linking were conducted:

- Gold linkage, in which name, address and other variables were used for linking
- Bronze linkage, in which Mesh Block and other variables (other than name and address information) were used for linking.

The aim of the Gold linked datasets is to serve as a benchmark against which the Bronze standards can be evaluated. The Bronze standards represent the type of linkage that can be conducted in future statistical studies between education data and the 2011 Census, since name and address are no longer available.

A research paper will be released later this year detailing the findings of the deterministic linkage exercises. Bronze and Statistical Linkage Key[4] (SLK) methods were used for the deterministic linkages.

The variables that were used in the Gold and Bronze probabilistic linkage exercises are shown in table 4.2.

3 Probabilistic linking compares records from two datasets using several variables common to both datasets and generates a single numerical measure of how well two particular records match. This allows ranking of all possible record pairs and assignment of the optimal link. Deterministic linking compares only record pairs that match exactly or almost exactly (e.g. age within one year) on a combination of variables, seeking unique matches wherever possible.

4 Many statistical linkage keys were devised for this study, to allow comparison between the various keys. The primary key that was used was SLK581. This is an alphanumeric match set key comprised of the second, third and fifth characters of a person's surname, the second and third letters of the person's first name, eight digits from the date of birth (DDMMYYYY) and one character representing the sex of the person (M or F), concatenated in that order. It must be noted that SLK581 is not a unique identifier given that if these personal detail components are the same for more than one person, they will have identical SLK values.

### 4.2 Variables used for Gold and Bronze standard linkage

| *Gold Standard* | | *Bronze Standard* | |
|---|---|---|---|
| *Blocking variables* | *Linking variables* | *Blocking variables* | *Linking variables* |
| First two letters of first name | First name | | |
| First two letters of surname | Surname | | |
| Age dummy (a) | Day of birth | Age dummy (a) | Day of birth |
| Day of birth | Month of birth | Age | Month of birth |
| Month of birth | Year of birth (±1) | | Age (±1) |
| Age | Age (±1) | | |
| Sex | | Sex | Sex |
| | Country of birth (2 digit) (b) | | Country of birth (2 digit) |
| | Main language spoken at home (2 digit) (c) | | Main language spoken at home (2 digit) |
| Statistical Area 1 2011 | Property name 2011 (c) | Mesh Block 2011 | |
| Postcode 2011 | Street number 2011 (d) | Statistical Area 1 2011 | |
| | Street name 2011 | | |
| | Suburb 2011 | Mesh Block 2010 | |
| | Postcode 2011 | | |
| | Street number 2010 | | |
| | Street name 2010 | | |
| | Suburb 2010 | | |
| | Postcode 2010 | | |
| | Street number 5 years ago (e) | | |
| | Street name 5 years ago (e) | | |
| | Suburb 5 years ago (e) | | |

(a) For the age dummy variable, all school enrolment records were set to 1, and Census records with age 3–25 years or imputed age were also set to 1.
(b) Not used for Tasmania.
(c) Not used for the Northern Territory.
(d) Not used for Queensland.
(e) Available on Census file only.
(f) Further information on the Australian Standard Geography Classification, including Mesh Block, Statistical Area 1 and Statistical Area 2 is available from ABS cat. no. 1270.0.55.001.

#### *4.2.1 Gold standard methodology*

Multiple passes (see Section 4.1) were run for the Gold standard to optimise the efficiency of linkage by limiting the number of comparisons in each pass (blocking) and to provide many opportunities for equivalent records to be compared (using different blocking variables for each pass). The overarching Gold blocking strategy is shown in table 4.3.

The blocking strategy used aimed to use a mix of blocking variables, starting with geographic location, then age variables. This strategy was repeated using broader, self-reported geography and Census address data from one year ago for linking variables. Tailored passes were then designed for each jurisdiction based on the agreement patterns that emerged with each combination of blocking and linking variables in order to maximise the linkage rates for each jurisdiction.

### 4.3 Gold standard blocking strategy

| *Pass no.* | *Blocking variables* | *Notes* |
|---|---|---|
| 1 (Geography) | Statistical Area 1<br>Age dummy(a) | |
| 2 (Age) | Day of birth<br>Month of birth<br>Exact age | |
| 3 (Geography) | Postcode 2011<br>Sex<br>Age dummy(a) | Only Queensland used Age dummy for this pass |
| 4 (Age) | Day of birth<br>Month of birth<br>Exact age | Linked using 2011 address from enrolments and address one year ago from Census |
| | TAILORED PASSES FOR EACH JURISDICTION | |
| | QUEENSLAND | |
| 5 (Name) | First two letters of first name & surname<br>Sex<br>Age dummy(a) | |
| 6 (Age / prior addresses) | Day of birth<br>Month of birth<br>Exact age | Linked using 2010 address from enrolments and address five years ago from Census |
| 7 (Age / prior addresses) | Day of birth<br>Month of birth<br>Exact age | Linked using 2011 address from enrolments and address one year ago from Census |
| | SOUTH AUSTRALIA | |
| 5 (Geography) | Postcode 2011<br>Sex<br>Age dummy(a) | Repeated Pass 4 with Age dummy and additional clerical review |
| | TASMANIA | |
| 5 (Age / prior addresses) | Day of birth<br>Month of birth<br>Exact age | Linked using 2010 address from enrolments and address one year ago from Census |
| | NORTHERN TERRITORY | |
| 5 (Name / prior addresses) | First two letters of first name & surname<br>Sex | Linked using 2011 address from enrolments and address one year ago from Census |
| 6 (Name / prior addresses) | First two letters of first name & surname<br>Sex | Linked using 2011 address from enrolments and address five years ago from Census |

(a) All enrolment records are set to 1, Census records with age 3–25 years or imputed age are set to 1.

Clerical review was performed on a random sample of record pairs after each pass, to determine:

- the upper and lower cut-off weights that were optimal for assigning as many links as possible without excessively decreasing linkage accuracy
- the need for further clerical review, and if needed, the amount of clerical review to be done.

A single cut-off weight was set if the sample showed this was adequate to assign a high proportion of links with high accuracy. In this case, no further clerical review would be performed and unlinked records proceeded to the next pass. However, if a single cut-off weight could not be set without compromising accuracy, then an upper and lower cut-off weight were used and clerical review was conducted for record pairs with weights between the two cut-offs.

In total, approximately five per cent of records linked were assigned through clerical review. Table 4.4 shows the proportion of records assigned automatically or by clerical review for each jurisdiction. A far higher proportion of clerical review was needed for the Northern Territory to maintain linkage accuracy and achieve a reasonable linkage rate.

**4.4 Gold standard links, by assignment type and jurisdiction**

| | *Assigned automatically* | | *Assigned by clerical review* | |
|---|---|---|---|---|
| *Jurisdiction* | *no.* | *%* | *no.* | *%* |
| Queensland | 496,467 | 99.5 | 2,434 | 0.5 |
| South Australia | 171,637 | 98.7 | 2,322 | 1.3 |
| Tasmania | 57,796 | 97.3 | 1,594 | 2.7 |
| Northern Territory | 23,189 | 86.2 | 3,721 | 13.8 |

Clerical review is a time and labour intensive element of data integration projects and it can be prone to errors (Solon and Bishop, 2009: Section 5.3.3). While it is critical to examine a selection of record pairs manually to assess the quality of the automated linkage process and prepare for the next pass, it is also important to optimise the resource load so as to achieve the best value for effort.

### 4.2.1 Bronze standard methodology

For Bronze linkage, there were fewer available blocking and linking variables (see table 4.2). For the Bronze high standard, five passes were run. For each pass, an evaluation of the agreement pattern between the datasets was used to decide on a single cut-off weight. All record pairs above the single cut-off weight were automatically linked, and all below it were not linked. As with the Gold standard, records that were not linked remained in the pool of possible links and could be assigned in later passes, when different blocking variables were used.

An additional two passes were run after the Bronze high standard was completed. These two passes blocked on geographic area, sex and age, and assigned all record pairs as links. While the linkage accuracy of the Bronze low standard is therefore lower, all the extra links assigned for this standard match on age, sex and small geographic area (Mesh Block or SA1). The purpose of the Bronze low is to increase the coverage of the linked dataset, by assigning a link for most students, while ensuring that the links assigned are demographically similar, if not equivalent records. The impact of this methodology is discussed in Section 5.3.

Clerical review was not used to assign individual record pairs in Bronze linkage for two reasons. First, there was little additional information available for review which could be matched across the school enrolment and Census datasets. Second, the blocking and linking variables were all numeric – clerical review was unlikely to detect transcription or other errors.

Table 4.5 shows the blocking strategy for the Bronze linkage. Passes using other geography levels, such as Statistical Area 2 and State, were attempted, but could not be completed due to the high number of possible matches and the increased number of false links. The overarching strategy was to begin by blocking using the smallest geographical level possible, and gradually broaden geography. These blocks were then to be repeated, linking on statistical geography from one year ago.

### 4.5 Bronze standard blocking strategy

| *Pass no.* | *Blocking variables* | *Notes* |
|---|---|---|
| | BRONZE (HIGH) | |
| 1 | Mesh Block<br>Age dummy(a) | |
| 2 | Statistical Area 1<br>Age dummy(a) | |
| 3 | Mesh Block<br>Age dummy(a) | 2011 enrolment geography linked to Census geography from one year ago |
| 4 | Mesh Block<br>Age dummy(a) | Tasmania only, tolerance of ±1 year on age |
| | BRONZE (LOW)(b) | |
| 6 | Mesh Block<br>Exact age<br>Sex | |
| 7 | Statistical Area 1<br>Exact age<br>Sex | |

(a) All enrolment records are set to 1, Census records with age 3–25 years or imputed age are set to 1.
(b) The Bronze (low) standard assigned all links with a positive total weight.

The Gold and Bronze probabilistic linkage exercises produced linked datasets where almost 9 out of 10 school enrolment records were linked to an equivalent Census record (see table 5.1). The following section evaluates the quality of the linked datasets for the four jurisdictions by comparing the findings of the Gold and Bronze probabilistic linkage methods.

# 5. EVALUATION OF THE LINKAGE

There are a number of ways to evaluate the quality of linked datasets. The following methods were used in the Education Quality Study and are described in this section of the paper:

- comparison of the expected number to the actual number of links between enrolment records and Census
- examination of the properties of enrolment records that were not linked to a Census record
- calculation of match-link rate and link accuracy of the different Bronze standard linkages compared with Gold
- assessment of the under- or over-representation of sub-groups in the Bronze datasets compared with the Gold.

## 5.1 Comparing expected number of links to actual number of links

Initially, it is important to consider how many records might reasonably be expected to link. Students on the school enrolment datasets might be missing from the Census dataset for several reasons:

- they are temporarily out of the country on Census night
- they are missed by the Census, thus contributing to the Census undercount
- they emigrated from Australia before the Census
- they have died since their enrolment at school, but before the Census.

The last two of these reasons are less likely for the student population than for the population as a whole because school students are generally young (the vast majority are aged between 4 and 20 years). Although direct estimation of the individual impact of each of these elements was not possible for this study, they are jointly taken into account in the calculation of Estimated Resident Population (ERP) from Census counts (ABS, 2013). Therefore, the difference between ERP and Census counts can be used to approximate the expected number of links possible for the Education Quality Study.

The first step in the estimation of how many school enrolment records might actually be available for linkage with the Census was to remove Residents Temporarily Overseas (RTOs) from the ERP. The ratio of Census counts to ERP was then applied to school enrolments to adjust the original number of students by the estimated proportion of people in each state who completed a Census form. This adjustment factor is an estimate only, in that there is a lag between the ERP estimate (30 June 2011) and Census night (8 August 2011).

Some demographic groups are more likely to be missed by the Census (ABS, 2011c). To ensure that the undercount adjustment factor was applied proportionately, for each state, the enrolments data was broken into sex (male / female) and 5 year age groups (from 0–55 years) and each age group was adjusted as follows:

$$\text{Enrolments (state, age group)} \times \frac{\text{Census counts (state, age group)}}{\text{ERP}}$$

The expected links were then summed for each state. Table 5.1 shows the total number and expected number of enrolment records available for linking. It also shows the linkage rates before and after adjusting for the expected number of links for the Gold and Bronze linkage standards, to demonstrate the impact of Census net undercount on linkage.

**5.1 Linkage rates, adjusted for expected links, by jurisdiction**

| | *Queensland* | *South Australia* | *Tasmania* | *Northern Territory* |
|---|---|---|---|---|
| | ENROLMENTS | | | |
| Students (no.) | 566,057 | 194,766 | 68,393 | 35,272 |
| Expected links (no.) | 560,916 | 193,706 | 67,157 | 33,306 |
| | GOLD LINKED RECORDS | | | |
| Number of students linked (no.) | 498,901 | 173,959 | 59,390 | 26,910 |
| Proportion of students linked (%) | | | | |
| Pre-adjustment | 88.1 | 89.3 | 86.8 | 76.3 |
| Post-adjustment | 88.9 | 89.8 | 88.4 | 80.8 |
| | BRONZE (HIGH) LINKED RECORDS | | | |
| Number of students linked (no.) | 357,673 | 119,962 | 47,317 | 18,101 |
| Proportion of students linked (%) | | | | |
| Pre-adjustment | 63.2 | 61.6 | 69.2 | 51.3 |
| Post-adjustment | 63.8 | 61.9 | 70.5 | 54.3 |
| | BRONZE (LOW) LINKED RECORDS | | | |
| Number of students linked (no.) | 461,779 | 157,599 | 59,700 | 26,592 |
| Proportion of students linked (%) | | | | |
| Pre-adjustment | 81.6 | 80.9 | 87.3 | 75.4 |
| Post-adjustment | 82.3 | 81.4 | 88.9 | 79.8 |

After adjusting for net undercount, almost nine out of ten students with an enrolment record were linked on the Gold datasets for Queensland, South Australia and Tasmania. For the Northern Territory, eight out of ten students were linked in the Gold dataset. Census net undercount had the largest impact in the Northern Territory, where approximately 5.6% of the students with a school enrolment record would not be expected to have an equivalent Census record. In Queensland and South Australia, net undercount has a proportionally smaller impact on the success of the linkage process, with less than 1% of school enrolment records not expected to link.

## 5.2 Enrolment records that were not linked

The main reasons for not linking students to the Census are:

- the corresponding Census record does not exist (this was discussed in 5.1)
- the student's data on either the enrolment record or the corresponding Census record is insufficient 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. As discussed in Section 3.3.4, missing, invalid and conflicting data is problematic for linkage, and occurs both on the Census and the school enrolments datasets. In the following sections, unlinked records are analysed from the perspective of school enrolments, since these datasets define the scope of the present study (see Section 3.2.1).

Table 5.2 contains details of linking variables with missing or invalid values in the non-linked school enrolment records by linkage standard. In particular, missing or invalid address data was common in the unlinked school enrolment records.

Note that name and address were not used directly as blocking or linking variables in the Bronze standards. However, missing or invalid address data frequently prevented records from being assigned to a Mesh Block, Statistical Area 1 and/or Statistical Area 2 geography level. Wherever possible, data was repaired to generate geography codes for students, for instance, by checking the alignment between geography variables such as state, suburb and postcode. In addition, where the student address information on a school enrolment record did not provide enough data to ascribe geographic codes, the school address was used where possible as a proxy at the SA2 and SA1 level. For this reason, there is no missing data for SA1 in the Northern Territory – each student record either had valid residential address information or else valid school address information.

## 5.2 Missing or invalid values in non-linked records, by linkage standard and jurisdiction

| *Linking variable* | *% Enrolment records* | *% Gold non-linked records* | *% Bronze (high) non-linked records* | *% Bronze (low) non-linked records* |
|---|---|---|---|---|
| | QUEENSLAND | | | |
| Students (no.) | 566,057 | 67,156 | 208,384 | 104,278 |
| First name | – | – | . . | . . |
| Surname | – | – | . . | . . |
| Day of birth | – | – | – | – |
| Month of birth | – | – | – | – |
| Year of birth | – | – | – | – |
| Sex | – | – | – | – |
| Country of birth | 10.5 | 21.4 | 21.3 | 32.2 |
| Main language spoken at home | 0.2 | 0.3 | 0.2 | 0.2 |
| Statistical Area 1 | 2.7 | 5.0 | 7.2 | 14.5 |
| Mesh Block | 5.1 | 8.7 | 13.0 | 23.4 |
| Street number | 14.1 | 21.3 | 25.3 | 36.8 |
| Street name | 2.3 | 3.7 | 5.7 | 10.1 |
| Suburb | – | 0.2 | 0.1 | 0.2 |
| Postcode | – | 0.2 | 0.1 | 0.2 |
| | SOUTH AUSTRALIA | | | |
| Students (no.) | 194,766 | 20,807 | 74,804 | 37,167 |
| First name | – | – | . . | . . |
| Surname | – | – | . . | . . |
| Day of birth | – | – | – | – |
| Month of birth | – | – | – | – |
| Year of birth | – | – | – | – |
| Sex | – | – | – | – |
| Country of birth | – | – | – | – |
| Main language spoken at home | 87.5 | 82.8 | 87.1 | 86.7 |
| Statistical Area 1 | 9.6 | 20.1 | 25.1 | 50.5 |
| Mesh Block | 11.3 | 21.7 | 27.5 | 52.7 |
| Street number | 12.3 | 22.4 | 27.2 | 50.1 |
| Street name | 8.6 | 19.6 | 21.9 | 42.9 |
| Suburb | 5.7 | 14.5 | 14.9 | 29.9 |
| Postcode | 5.7 | 14.5 | 14.8 | 29.9 |

## 5.2 Missing or invalid values in non-linked records, by linkage standard and jurisdiction *(continued)*

| *Linking variable* | *% Enrolment records* | *% Gold non-linked records* | *% Bronze (high) non-linked records* | *% Bronze (low) non-linked records* |
|---|---|---|---|---|
| | TASMANIA | | | |
| Students (no.) | 68,393 | 9,003 | 21,076 | 8,693 |
| First name | – | – | . . | . . |
| Surname | – | – | . . | . . |
| Day of birth | – | – | – | – |
| Month of birth | – | – | – | – |
| Year of birth | – | – | – | – |
| Sex | – | – | – | – |
| Country of birth | 8.2 | 13.1 | 11.0 | 10.3 |
| Main language spoken at home | 7.5 | 12.1 | 10.2 | 9.1 |
| Statistical Area 1 | 2.8 | 8.2 | 9.1 | 22.1 |
| Mesh Block | 3.4 | 9.1 | 10.1 | 23.3 |
| Street number | 2.6 | 8.1 | 7.6 | 17.3 |
| Street name | 2.0 | 6.9 | 6.3 | 14.8 |
| Suburb | – | 0.1 | 0.1 | 0.2 |
| Postcode | 0.2 | 1.3 | 0.7 | 1.6 |
| | NORTHERN TERRITORY | | | |
| Students (no.) | 35,272 | 8,362 | 17,171 | 8,680 |
| First name | – | – | . . | . . |
| Surname | 0.1 | 0.1 | . . | . . |
| Day of birth | – | – | – | – |
| Month of birth | – | – | – | – |
| Year of birth | – | – | – | – |
| Sex | – | – | – | – |
| Country of birth | 6.1 | 7.3 | 7.0 | 7.0 |
| Main language spoken at home | 19.9 | 17.9 | 18.3 | 17.4 |
| Statistical Area 1 | – | – | – | – |
| Mesh Block | 12.5 | 20.5 | 22.7 | 32.4 |
| Street number | 30.6 | 46.5 | 45.3 | 51.3 |
| Street name | 27.1 | 42.4 | 40.3 | 45.5 |
| Suburb | 1.0 | 2.2 | 1.7 | 2.1 |
| Postcode | 1.0 | 2.3 | 1.8 | 2.1 |

– nil or rounded to zero (including null cells)
. . not applicable

## 5.3 Link accuracy and match-link rate

Matches are defined as record pairs in which the two records relate to the same person. The Gold standard linkage is assumed to identify all possible matches and is used as a benchmark for evaluating the Bronze standard linkages.

Link accuracy and match-link rate are two measures of the quality of the Bronze linkage. Both measures are based on the number of links in the Bronze dataset that are correct as determined by the Gold linkage (i.e. matches). Link accuracy is the proportion of all Bronze links that are correct (matches). Match-link rate is the proportion of all matches that are present in the Bronze dataset.

The relationship between Bronze links and matches is shown in figure 5.3.

**5.3 Method of calculating link accuracy and match-link rate**

| | | Match status from Gold standard | | |
|---|---|---|---|---|
| | | Matches | Non-matches | |
| Link status from Bronze standard | Links | True links | False links | Total links |
| | Non-links | False non-links | True non-links | |
| | | Total matches | | |

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

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

Link accuracy and match-link rate for both Bronze linkage standards are shown in table 5.4. Generally, to increase coverage in the Bronze dataset, lower accuracy is tolerated and to increase accuracy, loss in coverage must be tolerated.

**5.4 Link accuracy and match-link rate, by linkage standard and jurisdiction**

| *Linkage standard* | *Proportion of students linked (%)* | *Link accuracy (%)* | *Match-link rate (%)* |
|---|---|---|---|
| | QUEENSLAND | | |
| Gold | 88.1 | . . | . . |
| Bronze (high) | 63.2 | 97.6 | 70.0 |
| Bronze (low) | 81.6 | 83.9 | 77.6 |
| | SOUTH AUSTRALIA | | |
| Gold | 89.3 | . . | . . |
| Bronze (high) | 61.6 | 97.8 | 67.4 |
| Bronze (low) | 80.9 | 84.2 | 76.3 |
| | TASMANIA | | |
| Gold | 86.8 | . . | . . |
| Bronze (high) | 69.2 | 97.8 | 77.9 |
| Bronze (low) | 87.3 | 86.8 | 87.2 |
| | NORTHERN TERRITORY | | |
| Gold | 76.3 | . . | . . |
| Bronze (high) | 51.3 | 93.2 | 62.7 |
| Bronze (low) | 75.4 | 69.8 | 69.0 |

. . not applicable

The linkage rate (proportion of students linked) is similar for the Gold and Bronze low datasets. Tasmania has a slightly higher linkage rate for the Bronze low dataset than for the Gold (87.3% compared with 86.8%). This is possible because record pairs assigned in the Bronze low, while being demographically similar, may not have been true links (and hence, may have been rejected in the Gold linkage). In the Bronze high dataset, greater stringency was applied to assigning record pairs with high accuracy – these pairs had to agree on most linking variables to be assigned as links. As a result, the link accuracy for the Bronze high datasets is very high, but the linkage and match-link rates are lower for all jurisdictions. Tasmania had the highest linkage rate for the Bronze high standard at 69.2% and the Northern Territory had the lowest linkage rate at 51.3%.

Linkage accuracy was above 90% for all jurisdictions for the Bronze high datasets. South Australia and Tasmania had the highest level of link accuracy for the Bronze high, both at 97.8%. The Northern Territory had the lowest at 93.2%. The pattern was similar for the Bronze low standard, with Tasmania attaining the highest level of link accuracy at 86.8%, and the Northern Territory the lowest at 69.8%.

For all jurisdictions, an increase in the match-link rate from the Bronze high standard to the Bronze low standard is accompanied by a decrease in link accuracy. This effect was most pronounced for the Northern Territory. From Bronze high to Bronze low, the match-link rate increased by only 6.3 percentage points (the smallest increase among all four jurisdictions) at a cost of a decrease of 23.4 percentage points in link accuracy (the largest among the jurisdictions). Conversely, Tasmania had the greatest match-link rate increase between the Bronze high and low standards (9.3 percentage points). This was accompanied by the lowest decrease in link accuracy, at 11.0 percentage points.

The implications of the relationship between link accuracy and match-link rate are that for the purposes of research and analysis, the researcher must choose whether accuracy or coverage is more important. For longitudinal analysis, accuracy may be paramount, while in cross-sectional research coverage may be more important. A previous Census linkage study (Bishop, 2009) has shown that, for some types of analysis, a Bronze low dataset may perform more like a Gold dataset than does a Bronze high dataset. They concluded that, in essence, it is preferable to link each person to someone with similar characteristics wherever possible than not to link at all. Accordingly, the Bronze low linkage strategy for the Census Education Quality Study was designed to match students with a similar Census record, so that even where a link is false (in that it does not appear on the Gold dataset), it is still a link to a person of the same age and sex in a proximate location (at worst, within the same Statistical Area Level 1).

In all jurisdictions, the proportion of records linked for Aboriginal and Torres Strait Islander students was lower than for non-Indigenous students. The disparity was greatest in the Northern Territory where the difference in the Gold standard linkage was almost 21 percentage points. The rate of linkage was the least different across Indigenous status in Tasmania (5.4 percentage point difference for the Gold standard linkage).

Across linkage methods, the rate of linkage was similar for the Gold and Bronze low linked datasets, and for Tasmania, the linkage rate was slightly higher on the Bronze low than the Gold for both Aboriginal and Torres Strait Islander students (83.6% compared with 82.5%) and non-Indigenous (88.0% compared with 87.9%) students. This is possible because the linkage accuracy was lower (some false links were accepted). The linkage rate for the Bronze high standard is much lower due to more stringent controls on link accuracy.

Tasmania recorded the highest linkage accuracy for Aboriginal and Torres Strait Islander students for the Bronze standards and the match-link rate for Tasmania was most similar across Indigenous status. The Northern Territory showed lower link accuracy for Aboriginal and Torres Strait Islander students than for non-Indigenous students (13.2 percentage point difference for the Bronze high and 30.0 percentage point difference for the Bronze low respectively). The Northern Territory also showed a much lower match-link rate than the other jurisdictions for Aboriginal and Torres Strait Islander students, for both the Bronze high and low standards of linkage.

The interaction of several factors (Indigenous status, remoteness and quality of the input data) are the cause of these results, with Tasmania having only a small area of remote geography and the best address data of the four jurisdictions. The Northern Territory, by comparison, has a considerable population in remote and very remote areas, has a much higher proportion of Aboriginal and Torres Strait Islander students and several data quality issues with the input data (for instance, more missing data on both the enrolments and Census datasets, see tables 3.7 and 3.8). Table 5.5 shows the linkage rates, link accuracy and match-link rate for all jurisdictions by linkage standard and Indigenous status.

### 5.5 Link accuracy and match-link rate, by Indigenous status, linkage standard and jurisdiction

| *Linkage standard* | *Proportion of students linked (%)* | *Link accuracy (%)* | *Match-link rate (%)* |
|---|---|---|---|
| | ABORIGINAL AND TORRES STRAIT ISLANDER STUDENTS | | |
| | QUEENSLAND | | |
| Gold | 78.9 | . . | . . |
| Bronze (high) | 53.1 | 95.9 | 64.5 |
| Bronze (low) | 75.2 | 74.4 | 70.9 |
| | SOUTH AUSTRALIA | | |
| Gold | 74.6 | . . | . . |
| Bronze (high) | 47.2 | 96.2 | 60.9 |
| Bronze (low) | 69.7 | 72.0 | 67.3 |
| | TASMANIA | | |
| Gold | 82.5 | . . | . . |
| Bronze (high) | 64.1 | 97.8 | 76.0 |
| Bronze (low) | 83.6 | 83.9 | 85.0 |
| | NORTHERN TERRITORY | | |
| Gold | 64.6 | . . | . . |
| Bronze (high) | 36.1 | 84.1 | 47.0 |
| Bronze (low) | 66.2 | 51.4 | 52.7 |
| | NON-INDIGENOUS STUDENTS | | |
| | QUEENSLAND | | |
| Gold | 89.0 | . . | . . |
| Bronze (high) | 64.1 | 97.8 | 70.4 |
| Bronze (low) | 82.2 | 84.7 | 78.2 |
| | SOUTH AUSTRALIA | | |
| Gold | 90.1 | . . | . . |
| Bronze (high) | 62.4 | 97.8 | 67.7 |
| Bronze (low) | 81.5 | 84.7 | 76.7 |
| | TASMANIA | | |
| Gold | 87.9 | . . | . . |
| Bronze (high) | 70.6 | 97.8 | 78.5 |
| Bronze (low) | 88.0 | 87.5 | 87.6 |
| | NORTHERN TERRITORY | | |
| Gold | 85.4 | . . | . . |
| Bronze (high) | 63.4 | 97.3 | 72.2 |
| Bronze (low) | 82.7 | 81.4 | 78.8 |

. . not applicable

## 5.4 Under- and over-representation of population sub-groups

School enrolment data contains demographic and socio-economic information about students and their parents / caregivers. In this section, the linkage methods are evaluated on the basis of how well different population groups were represented in the linked datasets, and checked for potential selection bias[5]. The tables below show the distribution of students by age, grade, remoteness, Indigenous status, country of birth, main language spoken at home and parental school attainment (as identified on the school enrolment files). The distribution for each linkage method is provided for comparison. The final section (5.4.8) looks at one of these characteristics, parental school attainment, from the perspective of the Census and compares this with the corresponding information from the enrolment files.

### *5.4.1 Age of students*

Table 5.6 shows the distribution of students by age on the enrolments records and the Gold and Bronze linked datasets. In all jurisdictions, the distribution of students in the three main age groups (6–9 years, 10–14 years, and 15–19 years) is consistent across the original enrolment dataset and the three linked datasets. Five year old students are included in the youngest age group in order to prevent the otherwise small group of students aged less than five years from being identifiable.

Students from all ages were linked on the Gold and Bronze low standard datasets. On the Bronze high standard dataset, comparisons between the enrolment records and the Census records where age was outside of the range 3–25 years were not undertaken. These Census records were excluded from the linkage in order to make the number of record comparisons practicable and to increase the link accuracy (see Section 4.2.1). Considering that students outside this age group are an extremely small proportion of the student population, and are not necessarily 'school students' (see Section 3.3.1), this is not likely to be detrimental to the coverage of the linked dataset.

As a result, in table 5.6, the proportion of students in the youngest (0–5 years) and oldest (20 years or over) age groups show a much smaller proportion of students on the Bronze high datasets than on the enrolment records. For instance, in South Australia, 4.1% of students on the enrolment records were 20 years or over, compared with less than 1% on the Bronze high linked dataset.

5 Selection bias is bias arising from the selection of non-representative survey populations to represent the whole population, or from processes in the collection which systemically favour particular population sub-groups.

## 5.6 Age, by linkage standard and jurisdiction

| | Enrolments | | Linkage standard: Gold | | Bronze (high) | | Bronze (low) | |
|---|---|---|---|---|---|---|---|---|
| | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| | | | | QUEENSLAND | | | | |
| 0–5 years | 41,404 | 7.3 | 36,928 | 7.4 | 28,046 | 7.8 | 37,129 | 8.0 |
| 6–9 years | 172,818 | 30.5 | 153,192 | 30.7 | 110,436 | 30.9 | 137,140 | 29.7 |
| 10–14 years | 212,300 | 37.5 | 189,529 | 38.0 | 140,661 | 39.3 | 181,786 | 39.4 |
| 15–19 years | 138,209 | 24.4 | 118,467 | 23.7 | 78,401 | 21.9 | 104,814 | 22.7 |
| 20 yrs or over | 1,326 | 0.2 | 785 | 0.2 | 129 | – | 910 | 0.2 |
| *Total* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | | SOUTH AUSTRALIA | | | | |
| 0–5 years | 12,177 | 6.3 | 10,303 | 5.9 | 4,242 | 3.5 | 5,604 | 3.6 |
| 6–9 years | 52,612 | 27.0 | 47,584 | 27.4 | 34,405 | 28.7 | 41,626 | 26.4 |
| 10–14 years | 67,160 | 34.5 | 61,360 | 35.3 | 46,026 | 38.4 | 57,974 | 36.8 |
| 15–19 years | 54,923 | 28.2 | 48,054 | 27.6 | 34,725 | 28.9 | 45,902 | 29.1 |
| 20 yrs or over | 7,894 | 4.1 | 6,658 | 3.8 | 564 | 0.5 | 6,493 | 4.1 |
| *Total* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | | TASMANIA | | | | |
| 0–5 years | 1,928 | 2.8 | 1,711 | 2.9 | 1,439 | 3.0 | 1,820 | 3.0 |
| 6–9 years | 18,705 | 27.3 | 16,705 | 28.1 | 13,531 | 28.6 | 15,851 | 26.6 |
| 10–14 years | 24,340 | 35.6 | 21,826 | 36.8 | 17,967 | 38.0 | 21,933 | 36.7 |
| 15–19 years | 21,177 | 31.0 | 17,569 | 29.6 | 14,141 | 29.9 | 18,257 | 30.6 |
| 20 yrs or over | 2,243 | 3.3 | 1,579 | 2.7 | 239 | 0.5 | 1,839 | 3.1 |
| *Total* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | | NORTHERN TERRITORY | | | | |
| 0–5 years | 2,241 | 6.4 | 1,756 | 6.5 | 1,213 | 6.7 | 1,922 | 7.2 |
| 6–9 years | 11,826 | 33.5 | 9,002 | 33.5 | 5,918 | 32.7 | 8,201 | 30.8 |
| 10–14 years | 13,153 | 37.3 | 9,969 | 37.0 | 6,765 | 37.4 | 10,009 | 37.6 |
| 15–19 years | 7,841 | 22.2 | 6,055 | 22.5 | 4,176 | 23.1 | 6,285 | 23.6 |
| 20 yrs or over | 211 | 0.6 | 128 | 0.5 | 29 | 0.2 | 175 | 0.7 |
| *Total* | *35,272* | *100.0* | *26,910* | *100.0* | *18,101* | *100.0* | *26,592* | *100.0* |

– nil or rounded to zero (including null cells)

### 5.4.2 Grade level

In addition to the linked datasets being representative of the enrolment files in terms of five-year age groups, students at each individual grade level were also well represented (table 5.7). As previously discussed (Section 5.4.1) on the Bronze high standard dataset, comparisons between the enrolment records and the Census records where age was outside of the range 3–25 years were excluded.

The proportion of students in Early childhood / Pre-school (Queensland only) is affected by this exclusion, with a smaller proportion on the Bronze high linked dataset. Also, the number of students in Grade 12 for all jurisdictions is lower on the Bronze high dataset than for school enrolments or the other linked datasets. For instance, in Tasmania, 7.7% of students on the enrolment records were in Grade 12, compared with 6.1% on the Bronze high linked dataset.

Note that table 5.7 includes only students who were enrolled in 2011.

## 5.7 Grade level, by linkage standard and jurisdiction, students enrolled in 2011

| | | | *Linkage standard* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Enrolments* | | *Gold* | | *Bronze High* | | *Bronze Low* | |
| | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| | QUEENSLAND | | | | | | | |
| Early childhood / pre-school | 2,664 | 0.5 | 2,350 | 0.5 | 1,715 | 0.5 | 2,444 | 0.6 |
| Pre-Grade 1 | 44,007 | 8.9 | 39,396 | 8.8 | 30,024 | 8.9 | 39,350 | 9.2 |
| Grade 1 | 42,165 | 8.5 | 37,725 | 8.5 | 28,484 | 8.4 | 37,045 | 8.7 |
| Grade 2 | 40,743 | 8.2 | 36,715 | 8.2 | 27,682 | 8.2 | 35,799 | 8.4 |
| Grade 3 | 39,331 | 8.0 | 35,328 | 7.9 | 26,390 | 7.8 | 33,145 | 7.8 |
| Grade 4 | 26,281 | 5.3 | 23,723 | 5.3 | 16,755 | 5.0 | 17,215 | 4.0 |
| Grade 5 | 39,810 | 8.1 | 36,020 | 8.1 | 25,579 | 7.6 | 31,158 | 7.3 |
| Grade 6 | 41,290 | 8.4 | 37,369 | 8.4 | 28,576 | 8.5 | 36,473 | 8.5 |
| Grade 7 | 40,292 | 8.1 | 36,404 | 8.2 | 27,745 | 8.2 | 35,383 | 8.3 |
| Grade 8 (secondary) | 36,320 | 7.3 | 32,845 | 7.4 | 25,477 | 7.5 | 32,264 | 7.6 |
| Grade 9 | 36,548 | 7.4 | 33,086 | 7.4 | 25,658 | 7.6 | 32,600 | 7.6 |
| Grade 10 | 37,743 | 7.6 | 33,863 | 7.6 | 26,192 | 7.8 | 33,423 | 7.8 |
| Grade 11 | 34,688 | 7.0 | 31,282 | 7.0 | 24,046 | 7.1 | 30,905 | 7.2 |
| Grade 12 | 28,998 | 5.9 | 26,484 | 5.9 | 20,619 | 6.1 | 26,440 | 6.2 |
| Ungraded | 3,557 | 0.7 | 3,203 | 0.7 | 2,529 | 0.7 | 3,180 | 0.7 |
| *Total* | *494,437* | *100.0* | *445,793* | *100.0* | *337,471* | *100.0* | *426,824* | *100.0* |
| | SOUTH AUSTRALIA | | | | | | | |
| Pre-Grade 1 | 16,033 | 9.5 | 13,811 | 9.0 | 6,984 | 6.5 | 9,038 | 6.6 |
| Grade 1 | 11,962 | 7.1 | 10,902 | 7.1 | 8,186 | 7.6 | 10,316 | 7.6 |
| Grade 2 | 12,284 | 7.3 | 11,246 | 7.4 | 8,363 | 7.8 | 10,578 | 7.8 |
| Grade 3 | 11,936 | 7.1 | 10,917 | 7.1 | 8,001 | 7.5 | 9,465 | 6.9 |
| Grade 4 | 12,216 | 7.2 | 11,260 | 7.4 | 7,664 | 7.1 | 8,328 | 6.1 |
| Grade 5 | 12,130 | 7.2 | 11,142 | 7.3 | 8,479 | 7.9 | 10,549 | 7.7 |
| Grade 6 | 12,383 | 7.3 | 11,435 | 7.5 | 8,569 | 8.0 | 10,755 | 7.9 |
| Grade 7 | 12,415 | 7.4 | 11,521 | 7.5 | 8,708 | 8.1 | 10,882 | 8.0 |
| Ungraded primary | 2,848 | 1.7 | 2,479 | 1.6 | 1,695 | 1.6 | 2,018 | 1.5 |
| Grade 8 (secondary) | 11,465 | 6.8 | 10,610 | 6.9 | 7,969 | 7.4 | 9,942 | 7.3 |
| Grade 9 | 11,710 | 6.9 | 10,764 | 7.0 | 8,006 | 7.5 | 10,091 | 7.4 |
| Grade 10 | 12,423 | 7.4 | 11,180 | 7.3 | 8,390 | 7.8 | 10,565 | 7.8 |
| Grade 11 | 14,436 | 8.6 | 12,784 | 8.4 | 8,297 | 7.7 | 11,865 | 8.7 |
| Grade 12 | 10,687 | 6.3 | 9,681 | 6.3 | 6,554 | 6.1 | 9,230 | 6.8 |
| Ungraded secondary | 3,576 | 2.1 | 3,125 | 2.0 | 1,446 | 1.3 | 2,610 | 1.9 |
| *Total* | *168,504* | *100.0* | *152,857* | *100.0* | *107,311* | *100.0* | *136,232* | *100.0* |

## 5.7 Grade level, by linkage standard and jurisdiction, students enrolled in 2011 *(continued)*

| | | | *Linkage standard* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Enrolments* | | *Gold* | | *Bronze High* | | *Bronze Low* | |
| | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| | | | | TASMANIA | | | | |
| Pre-Grade 1 | 4,763 | 8.0 | 4,241 | 8.0 | 3,492 | 8.1 | 4,361 | 8.3 |
| Grade 1 | 4,390 | 7.4 | 3,985 | 7.5 | 3,243 | 7.5 | 4,010 | 7.6 |
| Grade 2 | 4,420 | 7.4 | 3,995 | 7.5 | 3,312 | 7.7 | 3,985 | 7.6 |
| Grade 3 | 4,630 | 7.8 | 4,197 | 7.9 | 3,372 | 7.8 | 3,648 | 6.9 |
| Grade 4 | 4,461 | 7.5 | 4,064 | 7.6 | 3,247 | 7.6 | 3,704 | 7.0 |
| Grade 5 | 4,819 | 8.1 | 4,402 | 8.3 | 3,633 | 8.5 | 4,359 | 8.3 |
| Grade 6 | 4,840 | 8.1 | 4,406 | 8.3 | 3,635 | 8.5 | 4,407 | 8.4 |
| Grade 7 (secondary) | 4,263 | 7.2 | 3,885 | 7.3 | 3,179 | 7.4 | 3,872 | 7.4 |
| Grade 8 | 4,272 | 7.2 | 3,850 | 7.2 | 3,200 | 7.4 | 3,877 | 7.4 |
| Grade 9 | 4,519 | 7.6 | 4,037 | 7.6 | 3,300 | 7.7 | 4,083 | 7.8 |
| Grade 10 | 4,688 | 7.9 | 4,171 | 7.8 | 3,413 | 7.9 | 4,196 | 8.0 |
| Grade 11 | 4,860 | 8.2 | 4,144 | 7.8 | 3,323 | 7.7 | 4,176 | 7.9 |
| Grade 12 | 4,595 | 7.7 | 3,814 | 7.2 | 2,636 | 6.1 | 3,959 | 7.5 |
| *Total* | *59,520* | *100.0* | *53,191* | *100.0* | *42,985* | *100.0* | *52,637* | *100.0* |
| | | | | NORTHERN TERRITORY | | | | |
| Pre-Grade 1 | 2,796 | 9.6 | 2,179 | 9.6 | 1,497 | 9.6 | 2,347 | 10.6 |
| Grade 1 | 2,577 | 8.9 | 1,993 | 8.8 | 1,387 | 8.9 | 2,036 | 9.2 |
| Grade 2 | 2,542 | 8.8 | 1,957 | 8.7 | 1,339 | 8.6 | 1,982 | 8.9 |
| Grade 3 | 2,720 | 9.4 | 2,095 | 9.3 | 1,424 | 9.1 | 1,987 | 8.9 |
| Grade 4 | 2,658 | 9.2 | 2,075 | 9.2 | 1,257 | 8.1 | 1,388 | 6.2 |
| Grade 5 | 2,584 | 8.9 | 1,987 | 8.8 | 1,384 | 8.9 | 1,977 | 8.9 |
| Grade 6 | 2,470 | 8.5 | 1,896 | 8.4 | 1,322 | 8.5 | 1,878 | 8.5 |
| Grade 7 | 1,986 | 6.8 | 1,531 | 6.8 | 1,061 | 6.8 | 1,531 | 6.9 |
| Grade 8 (secondary) | 1,994 | 6.9 | 1,540 | 6.8 | 1,060 | 6.8 | 1,571 | 7.1 |
| Grade 9 | 1,802 | 6.2 | 1,414 | 6.3 | 988 | 6.3 | 1,456 | 6.6 |
| Grade 10 | 1,807 | 6.2 | 1,440 | 6.4 | 1,051 | 6.7 | 1,476 | 6.6 |
| Grade 11 | 1,758 | 6.1 | 1,406 | 6.2 | 972 | 6.2 | 1,426 | 6.4 |
| Grade 12 | 1,242 | 4.3 | 1,053 | 4.7 | 802 | 5.1 | 1,091 | 4.9 |
| Ungraded | 76 | 0.3 | 58 | 0.3 | 47 | 0.3 | 64 | 0.3 |
| *Total* | *29,012* | *100.0* | *22,624* | *100.0* | *15,591* | *100.0* | *22,210* | *100.0* |

### 5.4.3 Remoteness

Linkage was somewhat impeded for students from more remote areas. The proportion of records linked in more remote areas changes only slightly for each linkage method but was worst for the Bronze high standard, where geography was used extensively as a blocking variable.

Table 5.8 displays the distribution of students by remoteness areas and linkage standard for each jurisdiction. The remoteness area in Table 5.8 is based on the student's residential address on the enrolment record. A small proportion of students had a residential address on their enrolment record that was outside of the jurisdiction in which they were enrolled. These students, and students who did not have a valid residential address on their enrolment record/s, have been included in totals only.

For Aboriginal and Torres Strait Islander students, remoteness had a greater impact on the ability to link records, with the proportion of linked records for Aboriginal and Torres Strait Islander students decreasing substantially in the more remote areas. The quality of linking and findings from linkage for Aboriginal and Torres Strait Islander students will be examined more closely in a research paper to be released later this year (2013).

## 5.8 Remoteness, by linkage standard and jurisdiction

| | | | *Linkage standard* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Enrolments* | | *Gold* | | *Bronze (high)* | | *Bronze (low)* | |
| | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| | | | | QUEENSLAND | | | | |
| Major Cities | 327,297 | 57.8 | 289,270 | 58.0 | 212,791 | 59.5 | 271,798 | 58.9 |
| Inner Regional | 125,084 | 22.1 | 111,355 | 22.3 | 79,930 | 22.3 | 102,778 | 22.3 |
| Outer Regional | 92,853 | 16.4 | 80,558 | 16.1 | 54,619 | 15.3 | 72,158 | 15.6 |
| Remote | 10,528 | 1.9 | 8,944 | 1.8 | 5,362 | 1.5 | 7,628 | 1.7 |
| Very Remote | 9,262 | 1.6 | 7,916 | 1.6 | 4,419 | 1.2 | 6,533 | 1.4 |
| *Total(a)* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | | SOUTH AUSTRALIA | | | | |
| Major Cities | 129,906 | 66.7 | 118,864 | 68.3 | 84,535 | 70.5 | 111,177 | 70.5 |
| Inner Regional | 20,373 | 10.5 | 19,488 | 11.2 | 12,607 | 10.5 | 16,955 | 10.8 |
| Outer Regional | 35,429 | 18.2 | 27,342 | 15.7 | 17,750 | 14.8 | 22,785 | 14.5 |
| Remote | 6,329 | 3.2 | 6,115 | 3.5 | 3,921 | 3.3 | 5,006 | 3.2 |
| Very Remote | 2,477 | 1.3 | 1,937 | 1.1 | 1,033 | 0.9 | 1,476 | 0.9 |
| *Total(a)* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | | TASMANIA | | | | |
| Major Cities | . . | . . | . . | . . | . . | . . | . . | . . |
| Inner Regional | 42,893 | 62.7 | 37,194 | 62.6 | 30,039 | 63.5 | 38,024 | 63.7 |
| Outer Regional | 24,154 | 35.3 | 21,033 | 35.4 | 16,421 | 34.7 | 20,613 | 34.5 |
| Remote | 1,023 | 1.5 | 886 | 1.5 | 658 | 1.4 | 813 | 1.4 |
| Very Remote | 289 | 0.4 | 259 | 0.4 | 186 | 0.4 | 221 | 0.4 |
| *Total(a)* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | | NORTHERN TERRITORY | | | | |
| Major Cities | . . | . . | . . | . . | . . | . . | . . | . . |
| Inner Regional | . . | . . | . . | . . | . . | . . | . . | . . |
| Outer Regional | 18,772 | 53.2 | 15,469 | 57.5 | 11,573 | 63.9 | 15,356 | 57.7 |
| Remote | 6,122 | 17.4 | 4,606 | 17.1 | 2,852 | 15.8 | 4,309 | 16.2 |
| Very Remote | 10,340 | 29.3 | 6,798 | 25.3 | 3,655 | 20.2 | 6,894 | 25.9 |
| *Total(a)* | *35,272* | *100.0* | *26,910* | *100.0* | *18,101* | *100.0* | *26,592* | *100.0* |

. . not applicable

(a) Remoteness is based on the student's residential address on the enrolment record. Students who had a residential address on their enrolment record that was outside of the jurisdiction in which they were enrolled, and students without a valid residential address, have been included in totals only.

### 5.4.4 Aboriginal and Torres Strait Islander students

Indigenous status was available on both datasets (school enrolments and Census) but was not used as a blocking or linking variable in order to avoid any bias in subsequent analysis of the linked datasets by Indigenous status.

The proportionate distribution of students by Indigenous status is reasonably consistent between the school enrolment files and the three linkage methods. However, the actual number of Aboriginal and/or Torres Strait Islander students identified in the school enrolment files is generally higher than in any of the linked Census datasets (see Section 5.3 for more information about linkage accuracy and match-link rate by Indigenous status). Furthermore, there are fewer Aboriginal and/or Torres Strait Islander students represented on the Bronze high dataset than either the Gold or Bronze low standard, especially for the Northern Territory, as shown in table 5.9. As a result, the Bronze low standard dataset may be of greater utility for analysis of data for Aboriginal and/or Torres Strait Islander students.

## 5.9 Indigenous status, by linkage standard and jurisdiction

| | | | Linkage standard | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Enrolments | | Gold | | Bronze (high) | | Bronze (low) | |
| | no. | % | no. | % | no. | % | no. | % |
| | QUEENSLAND | | | | | | | |
| Indigenous | | | | | | | | |
| Aboriginal | 36,497 | 6.4 | 28,740 | 5.8 | 19,486 | 5.4 | 27,646 | 6.0 |
| Torres Strait Islander | 6,257 | 1.1 | 5,037 | 1.0 | 3,212 | 0.9 | 4,515 | 1.0 |
| Aboriginal and Torres Strait Islander | 5,514 | 1.0 | 4,307 | 0.9 | 2,932 | 0.8 | 4,118 | 0.9 |
| *Total* | *48,268* | *8.5* | *38,084* | *7.6* | *25,630* | *7.2* | *36,279* | *7.9* |
| Non-Indigenous | 517,459 | 91.4 | 460,563 | 92.3 | 331,880 | 92.8 | 425,245 | 92.1 |
| Not stated | 330 | 0.1 | 254 | 0.1 | 163 | – | 255 | 0.1 |
| *Total* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | SOUTH AUSTRALIA | | | | | | | |
| Indigenous | | | | | | | | |
| Aboriginal | 9,682 | 5.0 | 7,199 | 4.1 | 4,550 | 3.8 | 6,742 | 4.3 |
| Torres Strait Islander | 151 | 0.1 | 125 | 0.1 | 77 | 0.1 | 102 | 0.1 |
| Aboriginal and Torres Strait Islander | 78 | – | 66 | – | 48 | – | 63 | – |
| *Total* | *9,911* | *5.1* | *7,390* | *4.2* | *4,675* | *3.9* | *6,907* | *4.4* |
| Non-Indigenous | 184,855 | 94.9 | 166,569 | 95.8 | 115,287 | 96.1 | 150,692 | 95.6 |
| Not stated | – | – | – | – | – | – | – | – |
| *Total* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | TASMANIA | | | | | | | |
| Indigenous | | | | | | | | |
| Aboriginal | 4,946 | 7.2 | 4,102 | 6.9 | 3,187 | 6.7 | 4,141 | 6.9 |
| Torres Strait Islander | 175 | 0.3 | 135 | 0.2 | 109 | 0.2 | 149 | 0.2 |
| Aboriginal and Torres Strait Islander | 409 | 0.6 | 324 | 0.5 | 249 | 0.5 | 331 | 0.6 |
| *Total* | *5,530* | *8.1* | *4,561* | *7.7* | *3,545* | *7.5* | *4,621* | *7.7* |
| Non-Indigenous | 58,660 | 85.8 | 51,549 | 86.8 | 41,401 | 87.5 | 51,619 | 86.5 |
| Not stated | 4,203 | 6.1 | 3,280 | 5.5 | 2,371 | 5.0 | 3,460 | 5.8 |
| *Total* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | NORTHERN TERRITORY | | | | | | | |
| Indigenous | | | | | | | | |
| Aboriginal | 14,438 | 40.9 | 9,245 | 34.4 | 5,031 | 27.8 | 9,482 | 35.7 |
| Torres Strait Islander | 95 | 0.3 | 76 | 0.3 | 64 | 0.4 | 76 | 0.3 |
| Aboriginal and Torres Strait Islander | 968 | 2.7 | 694 | 2.6 | 496 | 2.7 | 698 | 2.6 |
| *Total* | *15,501* | *43.9* | *10,015* | *37.2* | *5,591* | *30.9* | *10,256* | *38.6* |
| Non-Indigenous | 19,606 | 55.6 | 16,752 | 62.3 | 12,424 | 68.6 | 16,212 | 61.0 |
| Not stated | 165 | 0.5 | 143 | 0.5 | 86 | 0.5 | 124 | 0.5 |
| *Total* | *35,272* | *100.0* | *26,910* | *100.0* | *18,101* | *100.0* | *26,592* | *100.0* |

– nil or rounded to zero (including null cells)

### 5.4.5 *Country of Birth*

Overall, students whose Country of Birth was not Australia had an increased likelihood of being linked as this extra information made it more likely that equivalent records would have a higher overall linkage weight. Students whose enrolment form indicated that they were born overseas were well represented on each of the linked datasets as the majority of records for students from diverse ethnic backgrounds were linked to their equivalent Census record. As shown in table 5.10, the distribution of records by the most common student countries of birth for each jurisdiction changes only slightly across linkage standards.

## 5.10 Country of Birth (selected countries), by linkage standard and jurisdiction

| | | | *Linkage standard* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Enrolments* | | *Gold* | | *Bronze (high)* | | *Bronze (low)* | |
| | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| | | | | QUEENSLAND | | | | |
| Australia | 445,212 | 78.7 | 400,049 | 80.2 | 300,110 | 83.9 | 382,291 | 82.8 |
| Overseas | | | | | | | | |
| New Zealand | 21,560 | 3.8 | 18,343 | 3.7 | 14,448 | 4.0 | 18,700 | 4.0 |
| England | 7,277 | 1.3 | 6,784 | 1.4 | 5,576 | 1.6 | 6,715 | 1.5 |
| South Africa | 3,020 | 0.5 | 2,876 | 0.6 | 2,292 | 0.6 | 2,741 | 0.6 |
| Philippines | 2,822 | 0.5 | 2,573 | 0.5 | 2,061 | 0.6 | 2,457 | 0.5 |
| India | 1,533 | 0.3 | 1,425 | 0.3 | 1,164 | 0.3 | 1,352 | 0.3 |
| Total Overseas | 61,636 | 10.9 | 54,027 | 10.8 | 42,793 | 12.0 | 53,851 | 11.7 |
| Missing / Not provided | 59,209 | 10.5 | 44,825 | 9.0 | 14,770 | 4.1 | 25,637 | 5.6 |
| *Total* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | | SOUTH AUSTRALIA | | | | |
| Australia | 169,497 | 87.0 | 152,693 | 87.8 | 105,852 | 88.2 | 137,548 | 87.3 |
| Overseas | | | | | | | | |
| England | 2,823 | 1.4 | 2,594 | 1.5 | 1,885 | 1.6 | 2,476 | 1.6 |
| China (exc. SARs and Taiwan) | 1,860 | 1.0 | 1,385 | 0.8 | 773 | 0.6 | 1,360 | 0.9 |
| India | 1,830 | 0.9 | 1,650 | 0.9 | 1,240 | 1.0 | 1,463 | 0.9 |
| U.K., Channel Is., Isle of Man, nfd | 1,533 | 0.8 | 1,385 | 0.8 | 1,024 | 0.9 | 1,315 | 0.8 |
| Philippines | 1,152 | 0.6 | 1,038 | 0.6 | 755 | 0.6 | 929 | 0.6 |
| Total Overseas | 25,269 | 13.0 | 21,266 | 12.2 | 14,110 | 11.8 | 20,051 | 12.7 |
| Missing / Not provided | – | – | – | – | – | – | – | – |
| *Total* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | | TASMANIA | | | | |
| Australia | 59,239 | 86.6 | 52,061 | 87.7 | 41,745 | 88.2 | 51,924 | 87.0 |
| Overseas | | | | | | | | |
| New Zealand | 291 | 0.4 | 249 | 0.4 | 202 | 0.4 | 246 | 0.4 |
| New Caledonia | 275 | 0.4 | 217 | 0.4 | 176 | 0.4 | 224 | 0.4 |
| U.K., Channel Is., Isle of Man, nfd | 128 | 0.2 | 112 | 0.2 | 87 | 0.2 | 119 | 0.2 |
| Western Sahara | 113 | 0.2 | 106 | 0.2 | 84 | 0.2 | 101 | 0.2 |
| Mongolia | 107 | 0.2 | 95 | 0.2 | 74 | 0.2 | 91 | 0.2 |
| Total Overseas | 3,556 | 5.2 | 2,908 | 4.9 | 2,285 | 4.8 | 3,074 | 5.1 |
| Missing / Not provided | 5,598 | 8.2 | 4,421 | 7.4 | 3,287 | 6.9 | 4,702 | 7.9 |
| *Total* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | | NORTHERN TERRITORY | | | | |
| Australia (inc. External Territories, nfd) | 30,115 | 85.4 | 22,949 | 85.3 | 15,256 | 84.3 | 22,600 | 85.0 |
| Overseas | | | | | | | | |
| Philippines | 595 | 1.7 | 514 | 1.9 | 437 | 2.4 | 510 | 1.9 |
| New Zealand | 247 | 0.7 | 178 | 0.7 | 120 | 0.7 | 187 | 0.7 |
| China (exc. SARs and Taiwan) | 156 | 0.4 | 132 | 0.5 | 106 | 0.6 | 134 | 0.5 |
| Thailand | 136 | 0.4 | 108 | 0.4 | 78 | 0.4 | 110 | 0.4 |
| India | 134 | 0.4 | 121 | 0.4 | 97 | 0.5 | 116 | 0.4 |
| Total Overseas | 2,990 | 8.5 | 2,406 | 8.9 | 1,888 | 10.4 | 2,431 | 9.1 |
| Missing / Not provided | 2,167 | 6.1 | 1,555 | 5.8 | 957 | 5.3 | 1,561 | 5.9 |
| *Total* | *35,272* | *100.0* | *26,910* | *100.0* | *18,101* | *100.0* | *26,592* | *100.0* |

– nil or rounded to zero (including null cells)

### 5.4.6 Main language spoken at home

Likewise, students whose Main language spoken at home was not English were well represented across each of the linked datasets. The majority of records for students from diverse ethnic backgrounds were linked to their equivalent Census records, as shown in table 5.11.

## 5.11 Main language spoken at home (selected languages), by linkage standard and jurisdiction

| | | | Linkage standard | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Enrolments | | Gold | | Bronze (high) | | Bronze (low) | |
| | no. | % | no. | % | no. | % | no. | % |
| | | | | QUEENSLAND | | | | |
| English | 528,498 | 93.4 | 465,909 | 93.4 | 332,144 | 92.9 | 429,837 | 93.1 |
| Vietnamese | 2,791 | 0.5 | 2,561 | 0.5 | 2,105 | 0.6 | 2,505 | 0.5 |
| Samoan | 2,632 | 0.5 | 2,228 | 0.4 | 1,862 | 0.5 | 2,267 | 0.5 |
| Mandarin | 2,168 | 0.4 | 1,970 | 0.4 | 1,477 | 0.4 | 1,872 | 0.4 |
| Yumplatok (Torres Strait Creole) | 1,812 | 0.3 | 1,416 | 0.3 | 780 | 0.2 | 1,235 | 0.3 |
| Australian Indigenous Languages, nfd | 1,665 | 0.3 | 1,364 | 0.3 | 812 | 0.2 | 1,308 | 0.3 |
| Other | 25,206 | 4.5 | 22,377 | 4.5 | 17,694 | 4.9 | 21,705 | 4.7 |
| Missing / Not provided | 1,285 | 0.2 | 1,076 | 0.2 | 799 | 0.2 | 1,050 | 0.2 |
| *Total* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | | SOUTH AUSTRALIA | | | | |
| English(a) | 4,656 | 2.4 | 4,133 | 2.4 | 3,092 | 2.6 | 3,907 | 2.5 |
| Vietnamese | 1,436 | 0.7 | 1,264 | 0.7 | 919 | 0.8 | 1,223 | 0.8 |
| Australian Indigenous Languages, nfd | 1,211 | 0.6 | 918 | 0.5 | 642 | 0.5 | 898 | 0.6 |
| Arabic | 895 | 0.5 | 732 | 0.4 | 482 | 0.4 | 679 | 0.4 |
| Chinese, nfd | 846 | 0.4 | 703 | 0.4 | 445 | 0.4 | 683 | 0.4 |
| Dari | 818 | 0.4 | 704 | 0.4 | 460 | 0.4 | 636 | 0.4 |
| Other | 14,495 | 7.4 | 12,331 | 7.1 | 8,698 | 7.3 | 11,402 | 7.2 |
| Missing / Not provided | 170,409 | 87.5 | 153,174 | 88.1 | 105,224 | 87.7 | 138,171 | 87.7 |
| *Total* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | | TASMANIA | | | | |
| English | 60,798 | 88.9 | 53,413 | 89.9 | 42,826 | 90.5 | 53,260 | 89.2 |
| Nepali | 204 | 0.3 | 194 | 0.3 | 159 | 0.3 | 192 | 0.3 |
| German | 135 | 0.2 | 93 | 0.2 | 65 | 0.1 | 106 | 0.2 |
| Arabic | 127 | 0.2 | 105 | 0.2 | 76 | 0.2 | 107 | 0.2 |
| Chinese, nfd | 126 | 0.2 | 74 | 0.1 | 56 | 0.1 | 101 | 0.2 |
| Korean | 114 | 0.2 | 64 | 0.1 | 46 | 0.1 | 90 | 0.2 |
| Other | 1,792 | 2.6 | 1,442 | 2.4 | 1,144 | 2.4 | 1,540 | 2.6 |
| Missing / Not provided | 5,097 | 7.5 | 4,005 | 6.7 | 2,945 | 6.2 | 4304 | 7.2 |
| *Total* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | | NORTHERN TERRITORY | | | | |
| English | 14,900 | 42.2 | 12,416 | 46.1 | 8,851 | 48.9 | 11,868 | 44.6 |
| Australian Indigenous Languages, nfd | 1,860 | 5.3 | 1,095 | 4.1 | 499 | 2.8 | 1,026 | 3.9 |
| Kriol | 1,471 | 4.2 | 995 | 3.7 | 524 | 2.9 | 947 | 3.6 |
| Djambarrpuyngu | 1,181 | 3.3 | 638 | 2.4 | 536 | 3.0 | 964 | 3.6 |
| Greek | 568 | 1.6 | 455 | 1.7 | 363 | 2.0 | 465 | 1.7 |
| Arrernte | 546 | 1.5 | 348 | 1.3 | 162 | 0.9 | 345 | 1.3 |
| Other | 7,739 | 21.9 | 5,449 | 20.2 | 3,298 | 18.2 | 5,478 | 20.6 |
| Missing / Not provided | 7,007 | 19.9 | 5,514 | 20.5 | 3,868 | 21.4 | 5,499 | 20.7 |
| *Total* | *35,272* | *100.0* | *26,910* | *100.0* | *18,101* | *100.0* | *26,592* | *100.0* |

– nil or rounded to zero (including null cells)

(a) In South Australia, the school enrolment form is intended to collect Main language other than English. As a result, Main language spoken at home is missing for most students and only a small proportion of students listed English.

### *5.4.7 Parent / caregiver school attainment – School enrolment form information*

Information about parents / caregivers on the school enrolment files includes country of birth, highest year of school completed, level of educational attainment and occupation. For the purpose of illustrating the quality of the linked datasets, only information on highest year of school completed is examined here. Additional parent / caregiver information and analyses are planned to be available in publications to be released later this year (2013).

The following tables show the school enrolments data on the highest year of school completed by parents / caregivers. Table 5.12 shows the data for Male parent / caregiver and 5.13 for Female parent / caregiver. Both tables show that where a student had no parent / caregiver information on their school enrolment file, they were less likely to be linked. As a consequence, the proportion of students with information on the schooling of their parent's / caregiver's (from their school enrolment file) is higher on the linked dataset than on the enrolments dataset.

Tables 5.12 and 5.13 should be read in conjunction with tables 5.14 and 5.15 (below), which display the corresponding data from the perspective of the Census.

## 5.12 Male parent / caregiver school educational attainment, by jurisdiction and linkage standard, school enrolment information

| | | | *Linkage standard* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Enrolments* | | *Gold* | | *Bronze (high)* | | *Bronze (low)* | |
| | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| | | | | QUEENSLAND | | | | |
| Year 12 or equivalent | 179,116 | 31.6 | 164,786 | 33.0 | 126,702 | 35.4 | 157,523 | 34.1 |
| Year 11 or equivalent | 36,285 | 6.4 | 32,823 | 6.6 | 25,027 | 7.0 | 31,504 | 6.8 |
| Year 10 or equivalent | 109,950 | 19.4 | 99,725 | 20.0 | 75,281 | 21.0 | 94,841 | 20.5 |
| Year 9 or equivalent or below | 24,617 | 4.3 | 21,559 | 4.3 | 16,310 | 4.6 | 20,754 | 4.5 |
| Not stated or unknown | 100,440 | 17.7 | 88,106 | 17.7 | 65,610 | 18.3 | 86,207 | 18.7 |
| Total students with a male parent | 450,408 | 79.6 | 406,999 | 81.6 | 308,930 | 86.4 | 390,829 | 84.6 |
| No Male parent information | 115,649 | 20.4 | 91,902 | 18.4 | 48,743 | 13.6 | 65,706 | 15.4 |
| *Total students* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | | SOUTH AUSTRALIA | | | | |
| Year 12 or equivalent | 60,513 | 31.1 | 54,888 | 31.6 | 39,460 | 32.9 | 49,870 | 31.6 |
| Year 11 or equivalent | 35,107 | 18.0 | 31,946 | 18.4 | 22,628 | 18.9 | 28,834 | 18.3 |
| Year 10 or equivalent | 26,677 | 13.7 | 24,149 | 13.9 | 17,219 | 14.4 | 21,903 | 13.9 |
| Year 9 or equivalent or below | 8,586 | 4.4 | 7,691 | 4.4 | 5,422 | 4.5 | 7,005 | 4.4 |
| Not stated or unknown | 33,235 | 17.1 | 28,960 | 16.6 | 19,124 | 15.9 | 26,231 | 16.6 |
| Total students with a male parent | 164,118 | 84.3 | 147,634 | 84.9 | 103,853 | 86.6 | 133,843 | 84.9 |
| No Male parent information | 30,648 | 15.7 | 26,325 | 15.1 | 16,109 | 13.4 | 23,756 | 15.1 |
| *Total students* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | | TASMANIA | | | | |
| Year 12 or equivalent | 14,813 | 21.7 | 13,375 | 22.5 | 10,986 | 23.2 | 13,311 | 22.3 |
| Year 11 or equivalent | 4,246 | 6.2 | 3,756 | 6.3 | 3,061 | 6.5 | 3,762 | 6.3 |
| Year 10 or equivalent | 24,100 | 35.2 | 21,769 | 36.7 | 17,857 | 37.7 | 21,378 | 35.8 |
| Year 9 or equivalent or below | 4,478 | 6.5 | 3,914 | 6.6 | 3,191 | 6.7 | 3,887 | 6.5 |
| Not stated or unknown | 3,934 | 5.8 | 3,181 | 5.4 | 2,409 | 5.1 | 3,318 | 5.6 |
| Total students with a male parent | 51,571 | 75.4 | 45,995 | 77.4 | 37,504 | 79.3 | 45,656 | 76.5 |
| No Male parent information | 16,822 | 24.6 | 13,395 | 22.6 | 9,813 | 20.7 | 14,044 | 23.5 |
| *Total students* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | | NORTHERN TERRITORY(a) | | | | |
| Year 12 or equivalent | 6,739 | 23.2 | 6,032 | 26.7 | 4,681 | 30.0 | 5,781 | 26.0 |
| Year 11 or equivalent | 2,449 | 8.4 | 2,120 | 9.4 | 1,627 | 10.4 | 2,077 | 9.4 |
| Year 10 or equivalent | 3,921 | 13.5 | 3,309 | 14.6 | 2,388 | 15.3 | 3,190 | 14.4 |
| Year 9 or equivalent or below | 2,071 | 7.1 | 1,585 | 7.0 | 1,017 | 6.5 | 1,495 | 6.7 |
| Not stated or unknown | 5,067 | 17.5 | 3,796 | 16.8 | 2,476 | 15.9 | 3,773 | 17.0 |
| Total students with a male parent | 20,247 | 69.8 | 16,842 | 74.4 | 12,189 | 78.2 | 16,316 | 73.5 |
| No Male parent information | 8,765 | 30.2 | 5,782 | 25.6 | 3,402 | 21.8 | 5,894 | 26.5 |
| *Total students* | *29,012* | *100.0* | *22,624* | *100.0* | *15,591* | *100.0* | *22,210* | *100.0* |

. . not applicable

(a) Students who were enrolled in 2010, but not in 2011, are excluded as parent sex data (derived from relationship data on the enrolments dataset) was not available for that year.

## 5.13 Female parent / caregiver school educational attainment, by jurisdiction and linkage standard, school enrolment information

| | | | *Linkage standard* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Enrolments* | | *Gold* | | *Bronze (high)* | | *Bronze (low)* | |
| | *no.* | *%* | *no.* | *%* | *no.* | *%* | *no.* | *%* |
| | | | | QUEENSLAND | | | | |
| Year 12 or equivalent | 230,334 | 40.7 | 210,596 | 42.2 | 161,159 | 45.1 | 201,022 | 43.5 |
| Year 11 or equivalent | 48,204 | 8.5 | 42,507 | 8.5 | 31,747 | 8.9 | 40,943 | 8.9 |
| Year 10 or equivalent | 109,964 | 19.4 | 98,034 | 19.6 | 73,779 | 20.6 | 94,121 | 20.4 |
| Year 9 or equivalent or below | 23,488 | 4.1 | 20,010 | 4.0 | 15,099 | 4.2 | 19,548 | 4.2 |
| Not stated or unknown | 107,224 | 18.9 | 92,955 | 18.6 | 68,784 | 19.2 | 91,533 | 19.8 |
| Total students with a Female parent | 519,214 | 91.7 | 464,102 | 93.0 | 350,568 | 98.0 | 447,167 | 96.8 |
| No Female parent information | 46,843 | 8.3 | 34,799 | 7.0 | 7,105 | 2.0 | 9,368 | 3.2 |
| *Total students* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | | SOUTH AUSTRALIA | | | | |
| Year 12 or equivalent | 73,261 | 37.6 | 66,337 | 38.1 | 47,196 | 39.3 | 59,961 | 38.0 |
| Year 11 or equivalent | 41,230 | 21.2 | 37,294 | 21.4 | 26,570 | 22.1 | 33,907 | 21.5 |
| Year 10 or equivalent | 28,321 | 14.5 | 25,402 | 14.6 | 17,931 | 14.9 | 23,079 | 14.6 |
| Year 9 or equivalent or below | 9,394 | 4.8 | 8,264 | 4.8 | 5,890 | 4.9 | 7,680 | 4.9 |
| Not stated or unknown | 31,194 | 16.0 | 26,833 | 15.4 | 16,551 | 13.8 | 23,977 | 15.2 |
| Total students with a Female parent | 183,400 | 94.2 | 164,130 | 94.3 | 114,138 | 95.1 | 148,604 | 94.3 |
| No Female parent information | 11,366 | 5.8 | 9,829 | 5.7 | 5,824 | 4.9 | 8,995 | 5.7 |
| *Total students* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | | TASMANIA | | | | |
| Year 12 or equivalent | 19,733 | 28.9 | 17,699 | 29.8 | 14,481 | 30.6 | 17,574 | 29.4 |
| Year 11 or equivalent | 7,546 | 11.0 | 6,745 | 11.4 | 5,421 | 11.5 | 6,601 | 11.1 |
| Year 10 or equivalent | 23,830 | 34.8 | 21,322 | 35.9 | 17,499 | 37.0 | 21,106 | 35.4 |
| Year 9 or equivalent or below | 3,607 | 5.3 | 3,004 | 5.1 | 2,463 | 5.2 | 3,060 | 5.1 |
| Not stated or unknown | 3,433 | 5.0 | 2,658 | 4.5 | 1,785 | 3.8 | 2,887 | 4.8 |
| Total students with a Female parent | 58,149 | 85.0 | 51,428 | 86.6 | 41,649 | 88.0 | 51,228 | 85.8 |
| No Female parent information | 10,244 | 15.0 | 7,962 | 13.4 | 5,668 | 12.0 | 8,472 | 14.2 |
| *Total students* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | | NORTHERN TERRITORY(a) | | | | |
| Year 12 or equivalent | 8,657 | 29.8 | 7,688 | 34.0 | 5,883 | 37.7 | 7,381 | 33.2 |
| Year 11 or equivalent | 3,245 | 11.2 | 2,693 | 11.9 | 1,980 | 12.7 | 2,639 | 11.9 |
| Year 10 or equivalent | 4,145 | 14.3 | 3,357 | 14.8 | 2,397 | 15.4 | 3,276 | 14.8 |
| Year 9 or equivalent or below | 2,805 | 9.7 | 1,978 | 8.7 | 1,220 | 7.8 | 1,885 | 8.5 |
| Not stated or unknown | 6,359 | 21.9 | 4,437 | 19.6 | 2,831 | 18.2 | 4,608 | 20.7 |
| Total students with a Female parent | 25,211 | 86.9 | 20,153 | 89.1 | 14,311 | 91.8 | 19,789 | 89.1 |
| No Female parent information | 3,801 | 13.1 | 2,471 | 10.9 | 1,280 | 8.2 | 2,421 | 10.9 |
| *Total students* | *29,012* | *100.0* | *22,624* | *100.0* | *15,591* | *100.0* | *22,210* | *100.0* |

. . not applicable
(a) Students who were enrolled in 2010, but not in 2011, are excluded as parent sex data (derived from relationship data on the enrolments dataset) was not available for that year.

### 5.4.8 Parent / caregiver school attainment – Census information

This final section briefly *compares* information obtained from school enrolments with the corresponding information as collected in the Census. It illustrates the potential for Census to supplement and enrich the socio-demographic information collected in administrative collections.

Parent / caregiver data on the enrolments form contained a high proportion of not stated for the school attainment variable. For those students for whom parent / caregiver data could be derived from the Census[6], the information on highest year of schooling contained a lower proportion of not stated data. That is, there is lower coverage of parent / caregiver data on the Census, but the available data is higher quality. This may have occurred for several reasons. First, there is a lower rate of non-response to the school educational attainment question on the Census than on the school enrolment files. Second, parents / caregivers could identify on the Census that they had not gone to school but this category was not identified on the school enrolment files. Third, if a student had information for only one parent / caregiver on their enrolment record, often the Census could deliver data for other parents / caregivers.

Note that the school enrolments data is also provided in the tables below for comparison purposes.

6 A complex process was undertaken to derive Census information for students' parents / caregivers (see the Explanatory Notes for more information about this derivation process).

## 5.14 Male parent / caregiver school educational attainment, by jurisdiction and linkage standard, Census information

| | | | Linkage standard | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Enrolments | | Gold | | Bronze (high) | | Bronze (low) | |
| | no. | % | no. | % | no. | % | no. | % |
| | | | QUEENSLAND | | | | | |
| Year 12 or equivalent | 179,116 | 31.6 | 166,109 | 33.3 | 123,144 | 34.4 | 161,175 | 34.9 |
| Year 11 or equivalent | 36,285 | 6.4 | 36,855 | 7.4 | 26,957 | 7.5 | 34,417 | 7.5 |
| Year 10 or equivalent | 109,950 | 19.4 | 113,419 | 22.7 | 82,068 | 22.9 | 104,448 | 22.6 |
| Year 9 or equivalent or below | 24,617 | 4.3 | 33,943 | 6.8 | 24,730 | 6.9 | 30,588 | 6.6 |
| Did not go to school | . . | . . | 1,823 | 0.4 | 1,397 | 0.4 | 1,670 | 0.4 |
| Not stated or unknown | 100,440 | 17.7 | 9,387 | 1.9 | 6,775 | 1.9 | 8,525 | 1.8 |
| Total students with a male parent | 450,408 | 79.6 | 361,536 | 72.5 | 265,071 | 74.1 | 340,823 | 73.8 |
| No Male parent information | 115,649 | 20.4 | 137,365 | 27.5 | 92,602 | 25.9 | 115,712 | 26.2 |
| *Total students* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | SOUTH AUSTRALIA | | | | | |
| Year 12 or equivalent | 60,513 | 31.1 | 53,004 | 30.5 | 38,907 | 32.4 | 49,844 | 31.6 |
| Year 11 or equivalent | 35,107 | 18.0 | 30,028 | 17.3 | 21,779 | 18.2 | 27,534 | 17.5 |
| Year 10 or equivalent | 26,677 | 13.7 | 23,223 | 13.3 | 16,578 | 13.8 | 20,861 | 13.2 |
| Year 9 or equivalent or below | 8,586 | 4.4 | 9,558 | 5.5 | 6,811 | 5.7 | 8,418 | 5.3 |
| Did not go to school | . . | . . | 938 | 0.5 | 696 | 0.6 | 875 | 0.6 |
| Not stated or unknown | 33,235 | 17.1 | 3,100 | 1.8 | 2,182 | 1.8 | 2,674 | 1.7 |
| Total students with a male parent | 164,118 | 84.3 | 119,851 | 68.9 | 86,953 | 72.5 | 110,206 | 69.9 |
| No Male parent information | 30,648 | 15.7 | 54,108 | 31.1 | 33,009 | 27.5 | 47,393 | 30.1 |
| *Total students* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | TASMANIA | | | | | |
| Year 12 or equivalent | 14,813 | 21.7 | 12,495 | 21.0 | 10,334 | 21.8 | 12,945 | 21.7 |
| Year 11 or equivalent | 4,246 | 6.2 | 3,748 | 6.3 | 2,995 | 6.3 | 3,738 | 6.3 |
| Year 10 or equivalent | 24,100 | 35.2 | 19,109 | 32.2 | 15,716 | 33.2 | 19,138 | 32.1 |
| Year 9 or equivalent or below | 4,478 | 6.5 | 3,975 | 6.7 | 3,234 | 6.8 | 3,875 | 6.5 |
| Did not go to school | . . | . . | 200 | 0.3 | 171 | 0.4 | 187 | 0.3 |
| Not stated or unknown | 3,934 | 5.8 | 1,073 | 1.8 | 871 | 1.8 | 1,038 | 1.7 |
| Total students with a male parent | 51,571 | 75.4 | 40,600 | 68.4 | 33,321 | 70.4 | 40,921 | 68.5 |
| No Male parent information | 16,822 | 24.6 | 18,790 | 31.6 | 13,996 | 29.6 | 18,779 | 31.5 |
| *Total students* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | NORTHERN TERRITORY(a) | | | | | |
| Year 12 or equivalent | 6,739 | 23.2 | 6,108 | 27.0 | 4,669 | 29.9 | 6,281 | 28.3 |
| Year 11 or equivalent | 2,449 | 8.4 | 2,660 | 11.8 | 1,951 | 12.5 | 2,580 | 11.6 |
| Year 10 or equivalent | 3,921 | 13.5 | 3,809 | 16.8 | 2,645 | 17.0 | 3,702 | 16.7 |
| Year 9 or equivalent or below | 2,071 | 7.1 | 2,500 | 11.1 | 1,493 | 9.6 | 2,300 | 10.4 |
| Did not go to school | . . | . . | 253 | 1.1 | 140 | 0.9 | 224 | 1.0 |
| Not stated or unknown | 5,067 | 17.5 | 631 | 2.8 | 355 | 2.3 | 545 | 2.5 |
| Total students with a male parent | 20,247 | 69.8 | 15,961 | 70.5 | 11,253 | 72.2 | 15,632 | 70.4 |
| No Male parent information | 8,765 | 30.2 | 6,663 | 29.5 | 4,338 | 27.8 | 6,578 | 29.6 |
| *Total students* | *29,012* | *100.0* | *22,624* | *100.0* | *15,591* | *100.0* | *22,210* | *100.0* |

. . not applicable

(a) Students who were enrolled in 2010, but not in 2011, are excluded as parent sex data (derived from relationship data on the enrolments dataset) was not available for that year.

## 5.15 Female parent / caregiver school educational attainment, by jurisdiction and linkage standard, Census information

| | | | Linkage standard | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Enrolments | | Gold | | Bronze (high) | | Bronze (low) | |
| | no. | % | no. | % | no. | % | no. | % |
| | | | QUEENSLAND | | | | | |
| Year 12 or equivalent | 230,334 | 40.7 | 228,665 | 45.8 | 169,971 | 47.5 | 220,583 | 47.8 |
| Year 11 or equivalent | 48,204 | 8.5 | 48,137 | 9.6 | 35,132 | 9.8 | 44,395 | 9.6 |
| Year 10 or equivalent | 109,964 | 19.4 | 124,658 | 25.0 | 90,413 | 25.3 | 114,329 | 24.8 |
| Year 9 or equivalent or below | 23,488 | 4.1 | 31,706 | 6.4 | 23,213 | 6.5 | 28,484 | 6.2 |
| Did not go to school | . . | . . | 2,565 | 0.5 | 1,994 | 0.6 | 2,357 | 0.5 |
| Not stated or unknown | 107,224 | 18.9 | 10,981 | 2.2 | 7,991 | 2.2 | 10,022 | 2.2 |
| Total students with a Female parent | 519,214 | 91.7 | 446,712 | 89.5 | 328,714 | 91.9 | 420,170 | 91.0 |
| No Female parent information | 46,843 | 8.3 | 52,189 | 10.5 | 28,959 | 8.1 | 41,609 | 9.0 |
| *Total students* | *566,057* | *100.0* | *498,901* | *100.0* | *357,673* | *100.0* | *461,779* | *100.0* |
| | | | SOUTH AUSTRALIA | | | | | |
| Year 12 or equivalent | 73,261 | 37.6 | 72,520 | 41.7 | 52,918 | 44.1 | 67,357 | 42.7 |
| Year 11 or equivalent | 41,230 | 21.2 | 35,904 | 20.6 | 26,361 | 22.0 | 32,972 | 20.9 |
| Year 10 or equivalent | 28,321 | 14.5 | 25,138 | 14.5 | 18,464 | 15.4 | 22,685 | 14.4 |
| Year 9 or equivalent or below | 9,394 | 4.8 | 9,674 | 5.6 | 6,985 | 5.8 | 8,510 | 5.4 |
| Did not go to school | . . | . . | 1,605 | 0.9 | 1,250 | 1.0 | 1,505 | 1.0 |
| Not stated or unknown | 31,194 | 16.0 | 3,808 | 2.2 | 2,716 | 2.3 | 3,365 | 2.1 |
| Total students with a Female parent | 183,400 | 94.2 | 148,649 | 85.5 | 108,694 | 90.6 | 136,394 | 86.5 |
| No Female parent information | 11,366 | 5.8 | 25,310 | 14.5 | 11,268 | 9.4 | 21,205 | 13.5 |
| *Total students* | *194,766* | *100.0* | *173,959* | *100.0* | *119,962* | *100.0* | *157,599* | *100.0* |
| | | | TASMANIA | | | | | |
| Year 12 or equivalent | 19,733 | 28.9 | 18,528 | 31.2 | 15,176 | 32.1 | 18,964 | 31.8 |
| Year 11 or equivalent | 7,546 | 11.0 | 6,285 | 10.6 | 5,180 | 10.9 | 6,224 | 10.4 |
| Year 10 or equivalent | 23,830 | 34.8 | 21,378 | 36.0 | 17,727 | 37.5 | 21,351 | 35.8 |
| Year 9 or equivalent or below | 3,607 | 5.3 | 3,238 | 5.5 | 2,670 | 5.6 | 3,187 | 5.3 |
| Did not go to school | . . | . . | 270 | 0.5 | 233 | 0.5 | 268 | 0.4 |
| Not stated or unknown | 3,433 | 5.0 | 1,308 | 2.2 | 1,070 | 2.3 | 1,280 | 2.1 |
| Total students with a Female parent | 58,149 | 85.0 | 51,007 | 85.9 | 42,056 | 88.9 | 51,274 | 85.9 |
| No Female parent information | 10,244 | 15.0 | 8,383 | 14.1 | 5,261 | 11.1 | 8,426 | 14.1 |
| *Total students* | *68,393* | *100.0* | *59,390* | *100.0* | *47,317* | *100.0* | *59,700* | *100.0* |
| | | | NORTHERN TERRITORY(a) | | | | | |
| Year 12 or equivalent | 8,657 | 29.8 | 8,154 | 36.0 | 6,235 | 40.0 | 8,373 | 37.7 |
| Year 11 or equivalent | 3,245 | 11.2 | 3,021 | 13.4 | 2,226 | 14.3 | 2,933 | 13.2 |
| Year 10 or equivalent | 4,145 | 14.3 | 4,159 | 18.4 | 2,866 | 18.4 | 3,904 | 17.6 |
| Year 9 or equivalent or below | 2,805 | 9.7 | 2,579 | 11.4 | 1,525 | 9.8 | 2,337 | 10.5 |
| Did not go to school | . . | . . | 292 | 1.3 | 173 | 1.1 | 260 | 1.2 |
| Not stated or unknown | 6,359 | 21.9 | 677 | 3.0 | 418 | 2.7 | 620 | 2.8 |
| Total students with a Female parent | 25,211 | 86.9 | 18,882 | 83.5 | 13,443 | 86.2 | 18,427 | 83.0 |
| No Female parent information | 3,801 | 13.1 | 3,742 | 16.5 | 2,148 | 13.8 | 3,783 | 17.0 |
| *Total students* | *29,012* | *100.0* | *22,624* | *100.0* | *15,591* | *100.0* | *22,210* | *100.0* |

. . not applicable

(a) Students who were enrolled in 2010, but not in 2011, are excluded as parent sex data (derived from relationship data on the enrolments dataset) was not available for that year.

# 6. CONCLUSIONS

The CDE Education Quality study has produced linked datasets that link a high proportion of government school enrolment records to the 2011 ABS Census of Population and Housing and accurately link equivalent records (that is the same individual from each dataset). While this is particularly the case for Gold linkage (using name and address), the Bronze datasets also display these qualities of coverage and accuracy. As a result, the linked datasets are highly representative of the enrolled school population.

As the tables in Section 5.4 clearly show, the majority of students in the enrolment records were present in the linked datasets, and were well represented across a range of demographic, geographic and social characteristics. Although lower than for the rest of the student population, the quality of linkage for Aboriginal and Torres Strait Islander students appears to be reasonable, especially for the Bronze low linkage, and will be further examined in a subsequent report to be released later this year (2013).

With the addition of Census data, the combined datasets contain enriched information about students, especially in terms of family characteristics such as family composition, parental education and household income. The school enrolment datasets provide better coverage of the school student population than does the Census (which has a relatively high proportion of missing data for type of educational institution). To gather the data that could be derived from the combined datasets from students and parents / caregivers through a detailed survey or extended enrolment form would require an impracticable degree of respondent burden and administrative work.

The use of name and address in the Gold linkage process and the ability to clerically review record pairs greatly increases the likelihood of identifying and linking equivalent records. Despite this, in all jurisdictions but most notably the Northern Territory, there remain a number of records that failed to be linked (Section 5.2).

If the records that were *not* linked are randomly distributed, it would be safe to assume that the linked datasets are representative of the entire student population. The analysis of missing data (Section 5.2) shows that the unlinked records contained a high proportion of missing or invalid address data or age data. Further analysis is required to examine whether these data quality issues are more likely to affect particular population sub-groups.

Two ways of reducing the gap between the entire student population and the linked records are to apply special methods targeting subgroups which are likely to be underrepresented, and to improve the input data. For instance, in the Gold data linkage, extensive clerical review was undertaken on the Northern Territory file (see table 4.4) to increase dramatically the number of records linked. Improvements in the collection of address and age information are possible and would lead to enhanced data quality for both the Census and for school enrolment records. For instance, the ABS is expanding the use of the Geocoded National Address File (G-NAF) in the Census to increase the proportion of records that can be successfully coded into geography levels (Mesh Block, Statistical Area 1, etc.) The high level of success of the linkage for Tasmania (especially in the Bronze standards) was almost solely attributable to the accuracy and completeness of school enrolment address data.

It is envisaged that future outputs from this study will also examine coverage and bias in the linked datasets to inform the selection of which linkage method is most appropriate for particular analytical purposes. For instance, the Bronze high standard has a much higher level of link accuracy than the Bronze low, which is likely to make it the preferred dataset for longitudinal analysis or data analysis requiring extensive disaggregation. Conversely, the Bronze low standard is more comprehensive, covering almost the same proportion of the student population as the Gold method. As a result, where cross-sectional or aggregate data is required, the Bronze low standard could be expected to deliver statistics that are representative of the whole student population.

It would appear therefore, that data linkage offers a viable alternative to new collections as a means of expanding the evidence base for education policy and research. The success of Bronze linkage for combining Census and school enrolment records indicates that probabilistic linkage without use of name and address could be used to combine other education collections, such as the Australian Early Development Index (AEDI) or national assessment programme (NAPLAN), with the Census or with each other. Since these collections draw demographic information from school enrolments, the quality of the linkage could be reasonably assumed to be similar to that examined in this study. Such projects could be undertaken in the near future and would continue the transformation of education and training data from discrete and somewhat fragmented individual collections to a more integrated research base of participation, attainment and socio-demographic information.

## REFERENCES


Australian Bureau of Statistics (2010) *Census Data Enhancement Project: An Update, Oct 2010*, cat. no. 2062.0, ABS, Canberra.

—— (2011a) *Australian Statistical Geography Standard (ASGS): Volume 1 – Main Structure and Greater Capital City Statistical Areas, July 2011*, cat. no. 1270.0.55.001, ABS, Canberra.

—— (2011b) *Census Dictionary, 2011*, cat. no. 2901.0, ABS, Canberra.

—— (2011c) *Measuring Net Undercount in the 2011 Population Census*, cat. no. 2940.0.55.001, ABS, Canberra.

—— (2013) *Australian Demographic Statistics, Sep 2012*, cat. no. 3101.0, ABS, Canberra.

Australian Curriculum, Assessment and Reporting Authority (2012) *Data Standards Manual,* ACARA, Sydney.

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.

Lloyd, J.E.V. and Hertzman, C. (2009) "From Kindergarten Rreadiness to Fourth-Grade Assessment: Longitudinal Analysis with Linked Population Data", *Social Science & Medicine*, 68, pp. 111–123.

Samuels, C. (2012) "Using the EM Algorithm to Estimate the Parameters of the Fellegi–Sunter Model for Data Linking", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.120, 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.

Wright, J.; Bishop, G. and Ayre, T. (2009) "Assessing the Quality of Linking Migrant Settlement Records to Census Data", *Methodology Research Papers*, cat. no. 1351.0.55.027, Australian Bureau of Statistics, Canberra.

# EXPLANATORY NOTES

## *1. Australian Standard Geographical Classification (ASGC)*

The ASGC provides a common framework of statistical geography which enables the production of statistics that are comparable and can be spatially integrated. To assign statistical geography, statistical units such as households are first assigned to a geographical area in one of the ASGC structures. Data collected from these statistical units are then compiled into ASGC defined geographic aggregations which, subject to confidentiality restrictions, are then available for publication. The geographic aggregations used for the purposes of this study are given below.

***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. Mesh Block is a useful linking variable when street address is not available.

***Statistical Area Level 1*** (SA1) is the second smallest geographic area defined in the Australian Statistical Geography standard (ASGS) after Mesh Block. The SA1 has been designed for use in the Census of Population and Housing as the smallest unit for the processing and release of Census data. SA1s are useful linking variables as they are still able to capture those who move within their local area without being so broad as to increase the possibility of matching different people who share similar characteristics, i.e. false links.

***Statistical Area Level 2*** (SA2) is an area defined in the ASGS, which consists of one or more whole Statistical Areas Level 1 (SA1s). Wherever possible SA2s are based on officially gazetted State suburbs and localities. In urban areas, SA2s largely conform to whole suburbs and combinations of whole suburbs, while in rural areas they define functional zones of social and economic links. This level is broad enough to capture the majority of matching pairs, where geocoding to the locality (town or suburb) has been reasonably accurate.

### *2. Derivation of Census information for students' parents / caregivers*

The process to derive students' parent / caregiver information from the Census began by selecting Census records that were likely to be records of parents / caregivers. This was done on the basis of family type in the dwelling, person age, sex and relationships within dwellings. From this subset, using the relationship between persons in the dwellings, parent / caregiver records that were connected to the linked student were selected. Only four parents / caregivers were selected – the student's natural mother and father, and the students step-mother and step-father. As a result, for students living with parents / caregivers who are in a same-sex relationship, only one parent / caregiver is selected. From this information, one overall male and female parent was selected – the natural parent where available, followed by step-parent information if natural parent information was not available.

## 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. |
| *LIBRARY* | A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries. |

### 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, or purchase a hard copy publication. 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 |