Australian Bureau of Statistics

1351.0.55.061

Research Paper

Assessing the Quality of Linking Temporary Visa Holders Administrative Data to the 2011 Census

New Issue

Research Paper

Assessing the Quality of Linking Temporary Visa Holders Administrative Data to the 2011 Census

National Migrant Statistics Unit

Population & Social Statistics Division

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 20 MAR 2017

ABS Catalogue no. 1351.0.55.061

© Commonwealth of Australia 2017

This work is copyright. Apart from any use as permitted under the *Copyright Act 1968*, no part may be reproduced by any process without prior written permission from the Commonwealth. Requests and inquiries concerning reproduction and rights in this publication should be addressed to The Manager, Intermediary Management, Australian Bureau of Statistics, Locked Bag 10, Belconnen ACT 2616, by telephone (02) 6252 6998, fax (02) 6252 7102, or email <intermediary.management@abs.gov.au>.

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Jenny Dobak, National Migrant Statistics Unit, on Adelaide (08) 8237 7317.

ASSESSING THE QUALITY OF LINKING TEMPORARY VISA HOLDERS ADMINISTRATIVE DATA TO THE 2011 CENSUS

National Migrant Statistics Unit
Australian Bureau of Statistics

EXECUTIVE SUMMARY

Temporary entrants are an important and growing part of the Australian population, with the latest available figures from the Department of Immigration and Border Protection (DIBP) showing close to 1.2 million temporary visa holders in the country as at 30 June 2016. This includes over 400,000 international students and approximately 170,000 skilled worker visa holders.

International students have a significant impact on the Australian economy through their contribution to the education export industry, while the availability of temporary skilled workers enables businesses to access international labour markets for skills and technical expertise. Increasingly, temporary entry is becoming a pathway to permanent residency. As such, the policy importance of this group is growing, and data is needed to provide a sound evidence base to adequately shape and evaluate relevant immigration and settlement policy. Data integration presents the opportunity to help meet this need for more information on temporary entrants.

In 2014, the National Migrant Statistics Unit of the Australian Bureau of Statistics (ABS) investigated the feasibility of probabilistically linking records from the 2011 Census of Population and Housing to administrative data on temporary visa holders held by DIBP. The linkage was deemed feasible and results were published in the research paper *Assessing the Suitability of Temporary Migrants Administrative Data for Data Integration* (ABS cat. no. 1351.0.55.053). Following on from this, the ABS has proceeded to link the temporary visa holders' records to the 2011 Census records.

An innovative linking strategy was developed to link the Temporary Visa Holders (TVH) dataset with the 2011 Census. In addition to standard geographic and demographic information, this strategy exploited comparable employment and education variables to improve the likelihood of finding unique links amongst a population characterised by a high degree of homogeneity. A dwelling indicator was also utilised to assist in linking secondary applicants.

At the completion of the linkage process, more than 240,000 of the 513,000 TVH records had been linked to a Census record to create the Australian Census and Temporary Entrants Integrated Dataset. This equates to a linkage rate of 48%. Given the characteristics and temporary nature of this population of interest, this is considered a good result.

The two main reasons for unlinked TVH records were found to be that either the person was not recorded in the Census or that the quality of information on the TVH and/or Census datasets was insufficient to adequately link the record. It is likely that factors such as language barriers and uncertainty about whether to participate in the Census may have contributed to a large number of temporary entrants not being included in the Census, or providing insufficient or inaccurate information.

The linked dataset is characterised by a degree of under and over-representation of some migrant subpopulations within the international student and skilled worker temporary entrant populations. For example, temporary visa holders from China and those aged between 18 and 25 years were found to be particularly under-represented. Hence, inferences made from the linked dataset about these subpopulations may be biased if appropriate adjustments for the unlinked components are not made.
As such, a weighting strategy to account for under- and over-representation of key subgroups in the dataset and support appropriate statistical analyses has been developed as part of this study.

Information contained on the TVH dataset was generally of high quality across most variables, with the exception of address information. The quality of address information varied considerably and in some instances was missing. For international students in particular, many of whom share quite generic characteristics, detailed geographic information gleaned from their address can be especially important in finding high quality, unique links. A combination of missing and poor quality geographic information can seriously affect linkage outcomes, especially in the absence of a unique identifier common to both datasets. Name information was also not used in the study. In the future, however, the potential for using anonymised name information as an additional linking variable could mitigate the issues caused by less than optimal address information and offer substantial improvements in linkage outcomes.

This paper concludes that with informed use, the linked dataset can be used for analysing the social and economic characteristics of temporary entrants by their visa and Census characteristics for the very first time. Two case studies are presented to demonstrate this. A dataset containing information representative of 240,000 temporary migrants is a significant step forward in filling a data gap that is well recognised. This dataset has the potential to inform a number of policy and research questions including the economic contributions of temporary visa holders, their geographic distribution, and the extent to which temporary workers are in jobs commensurate with their skills. Following on from these promising results, it is recommended that the linkage be repeated utilising the 2016 Census, and that the resulting linked and confidentialised dataset be made available for research and analytical purposes.

ACKNOWLEDGEMENTS

This paper was prepared by the ABS' National Migrants Statistics Unit with assistance from the ABS Data Integration, Access and Confidentiality Methodology Unit.
The National Migrant Statistics Unit work program is jointly funded by the ABS, DIBP and the Department of Social Services.

The study 'Assessing the Quality of Linking Temporary Visa Holders Administrative Data to the 2011 Census' was funded by DIBP and the results are based, in part, on administrative immigration data collected by DIBP. Any discussion on data limitations in this paper is in the context of using the data for statistical integration purposes, and is not related to the ability of the data to support DIBP's core operational requirements.

The confidentiality of data is protected by legislation. The *Census and Statistics Act 1905* require that all personal information collected by the ABS remain confidential. This Act ensures that data collected by the ABS cannot be released in a manner likely to enable an individual to be identified.

The *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes* (National Statistical Service, 2010) were adhered to in completing this linkage, further protecting the privacy of individuals.

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ACTEID	Australian Census and Temporary Entrants Integrated Dataset
ANZSCO	Australian and New Zealand Standard Classification of Occupations
ANZSIC	Australian and New Zealand Standard Industrial Classification
ASGS	Australian Statistical Geography Standard
COB	Country of Birth
BDAY	Day of Birth
BYEAR	Year of Birth
DIBP	Department of Immigration and Border Protection
EDU	Educational attainment
ELICOS	English Language Intensive Course for Overseas Students
ICSE	Integrated Client Services Environment
INC	Income
IND	Industry
IRIS	Immigration Records Information System
LEV	Level of educational institution being attended
MB	ASGS Mesh block
MSTA	Marital Status
PES	Post Enumeration Survey
PCODE	Postcode
SA1	ASGS Statistical Area Level 1
SA2	ASGS Statistical Area Level 2
TVH	Temporary Visa Holders
YOA	Year of Arrival

CONTENTS

APPENDIX

ASSESSING THE QUALITY OF LINKING TEMPORARY VISA HOLDERS ADMINISTRATIVE DATA TO THE 2011 CENSUS

National Migrant Statistics Unit
Australian Bureau of Statistics

ABSTRACT

In 2014, the Australian Bureau of Statistics was provided with a dataset of temporary visa holders in Australia as at 31 July 2011 by the Department of Immigration and Border Protection to assess its suitability for integration with the 2011 Census of Population and Housing. In 2015–16, the ABS proceeded to probabilistically link the records of international students and temporary skilled workers in Australia with person-level information from the 2011 Census. The resultant dataset, the Australian Census and Temporary Entrants Integrated Dataset, promises to contribute significantly to a better understanding of the socio-economic characteristics and geographic distribution of these two temporary entrant populations, leading to more informed and targeted policy development and evaluation. This paper describes the creation of the linked dataset and provides an assessment of its quality.

1. INTRODUCTION

Since 2011, the Australian Bureau of Statistics (ABS) has used statistical data integration to combine a growing number of administrative datasets with the Census of Population and Housing (Census). The creation of the Australian Census and Migrants Integrated Dataset, which successfully linked the Australian Government's administrative information on permanent migrants with the 2011 Census, provided the foundation for the exploratory work to link administrative data on temporary entrants to the Census. Temporary entrants are an important and growing part of the Australian population, with the latest available figures from the Department of Immigration and Border Protection (DIBP) showing close to 1.2 million temporary visa holders in the country as at 30 June 2016 (DIBP, 2016a).

In 2014, DIBP provided the ABS with access to the Temporary Visa Holders (TVH) administrative dataset containing information on international students and Temporary work (skilled) (subclass 457) visa holders, for the purposes of assessing whether it would be feasible to link the dataset to the 2011 Census. Utilising an experimental linkage simulation tool that simulates the probabilistic linking process and provides a report on the likely feasibility of linking two datasets prior to linkage, the ABS concluded that it would be feasible to link the TVH dataset with the 2011 Census. Further details are contained in the research paper *Assessing the Suitability of Temporary Migrants Administrative Data for Data Integration* (ABS, 2014).

Following on from this feasibility study, the ABS proceeded with the linkage of these two datasets, and has conducted an evaluation of the quality of the resulting linked dataset. Due to the lack of a common identifier on the two datasets, and Census name and address data being unavailable for use in linking having been destroyed following the completion of processing of the 2011 Census, probabilistic linking methods were employed, whereby linkage is based on the level of overall agreement on a set of variables common to both datasets.

The resulting linked dataset, known as the Australian Census and Temporary Entrants Integrated Dataset (ACTEID), will contribute to a better understanding of the socio-economic characteristics and geographic distribution of temporary entrants (international students and Temporary work (skilled) (subclass 457) visa holders). This will allow Governments to deliver more targeted migration policy development and evaluation.

This paper provides:

- An overview of the 2014 Feasibility Study (Section 2);
- Information on the datasets being linked (Section 3);
- A description of the linking methodology (Section 4);
- An evaluation of the quality of the resulting linked dataset (Section 5);
- A discussion on the use of a weighting strategy to address the under- and over-representation of subpopulations in the linked dataset (Section 6);
- A brief look at some potential applications for the linked dataset (Section 7); and
- Some recommendations for future linkages involving the source datasets (Section 8).

2. THE 2014 FEASIBILITY STUDY

In 2014, the ABS conducted a study which examined the suitability of administrative data on temporary visa holders for data integration. For the purposes of that study, the TVH dataset comprised of data on international students and Temporary work (skilled) (subclass 457) visa holders in Australia on 31 July 2011. The study investigated the quality of the administrative data and assessed its suitability for probabilistic linking with the 2011 Census of Population and Housing. For more information on the Feasibility Study, see the research paper *Assessing the Suitability of Temporary Migrants Administrative Data for Data Integration* (ABS, 2014).

The characteristics of the data items provided for the study were assessed for completeness and any anomalies. The feasibility of linking with the 2011 Census was assessed with assistance from the Data Integration, Access and Confidentiality Methodology section of the ABS, using an experimental linkage simulation tool. This tool simulates the probabilistic linking process and provides a diagnostic report enabling the feasibility of linking two datasets to be assessed prior to actual linkage.

The study concluded that linking the TVH data to the 2011 Census was feasible and could result in a maximum link rate of about 70% with 98% precision. Despite a number of not unexpected quality issues with the data, the study suggested that all the required elements were in place to produce a useful dataset for analysis. However, in order to improve the link rate, and provide more reliable and detailed estimates, more work was required in order to:

- assess and improve data quality, and
- address any issues that arise from the under-representation of certain subgroups in the linked dataset to effectively account for any systematic bias.

Following on from this study, the ABS conducted the actual linkage of these two datasets in 2015–16, with the continued support of the DIBP who supplied the data. The remainder of this paper focusses on this linkage.

3. SOURCES OF DATA FOR THE AUSTRALIAN CENSUS AND TEMPORARY MIGRANTS INTEGRATED DATASET (ACTEID)

This section provides an overview of the two datasets that were linked to create the Australian Census and Temporary Entrants Integrated Dataset (ACTEID) – the Temporary Visa Holders dataset and the 2011 Australian Census of Population and Housing.

3.1 Temporary Visa Holders dataset

The TVH dataset is compiled by DIBP from various departmental administrative systems and a number of external sources. It contains records of all persons in Australia holding an international student visa or a Temporary work (skilled) (subclass 457) visa as at 31 July 2011. While data is available from DIBP for a number of other temporary visas, this data was not included in this study because the number of visas issued are quite small when compared with international students and 457 visa holders. Whilst it is acknowledged that the numbers of people on Working Holiday visas in Australia is substantial, the majority of these people would not participate in the Census as they are not considered to be usual residents (due to the short term nature of their visas). The reference date of 31 July 2011 was selected as it was closest available to the Census date of 9 August 2011.

The TVH dataset is comprised of four extracts, which each contain unique record identifiers that enable the extracts to be combined (see Section 3.1.5).

The consolidated TVH dataset contained 513,184 records, which included core demographic and visa characteristics, along with address, education and employment information where available. Further detail on the four component extracts and the issues associated with bringing them together is provided below.

3.1.1 Visa Holders extract

This Visa Holders extract contained 513,184 records consisting of demographic information such as sex, date of birth, marital status and country of birth. The extract also contained visa information including visa subclass and applicant type (primary or secondary). Primary applicants must satisfy the criteria for the grant of a visa under the *Migration Regulations Act, 1994*, while secondary applicants are members of the family unit of the primary applicant (e.g. their spouse, partner, child or other relative).

The information on the Visa Holders extract was obtained from the following source systems:

- Offshore visa applications are processed using the Immigration Records Information System (IRIS); and
- Onshore visa applications are processed using the Integrated Client Services Environment (ICSE).

3.1.2 Client Address extract

The Client Address extract contained 473,697 records that provide various forms of address information pertaining to the visa holder. Recorded addresses may be residential, business, postal or other, and may be located in Australia or overseas. For some visa holders, multiple addresses, and/or a combination of address types were provided.

A number of issues with the address information on this extract were identified in the aforementioned Feasibility Study, and are summarised below:

- Address details are based on clients recorded in ICSE only. There is no opportunity to record addresses in the IRIS source system. Consequently, 11% of the Visa Holders did not have a corresponding client address record.
- The address data may be an onshore or offshore (overseas) address. An offshore address cannot be linked to a Census record as Census records have a place of usual residence within Australia.
- The address data provided may not be that of the temporary visa holder (e.g. a Migration Agent's address).
- There were several instances of the same address being reported multiple times. The most commonly reported residential addresses appeared to be student accommodation or the addresses of Migration Agents.
- The address data may be missing, incomplete or not up to date.

Subsequent investigation identified that 15.3% of primary applicants (14% of international students and 20% of temporary workers) had not provided an Australian address that would assist in the record linkage.

3.1.3 Student Confirmation of Enrolment extract

The Student Confirmation of Enrolment extract (the 'Student extract') contained 324,772 records for those temporary entrants in Australia on a Subclass 570-576 Visa, hereafter referred to as 'Students'. It pertains to primary applicants only, and provides information including the level of education currently being undertaken by the applicant, as well as the name of the institution at which the applicant is enrolled. Visa subclasses 570-576 enable applicants to stay in Australia to study for the duration of their course, as outlined below:

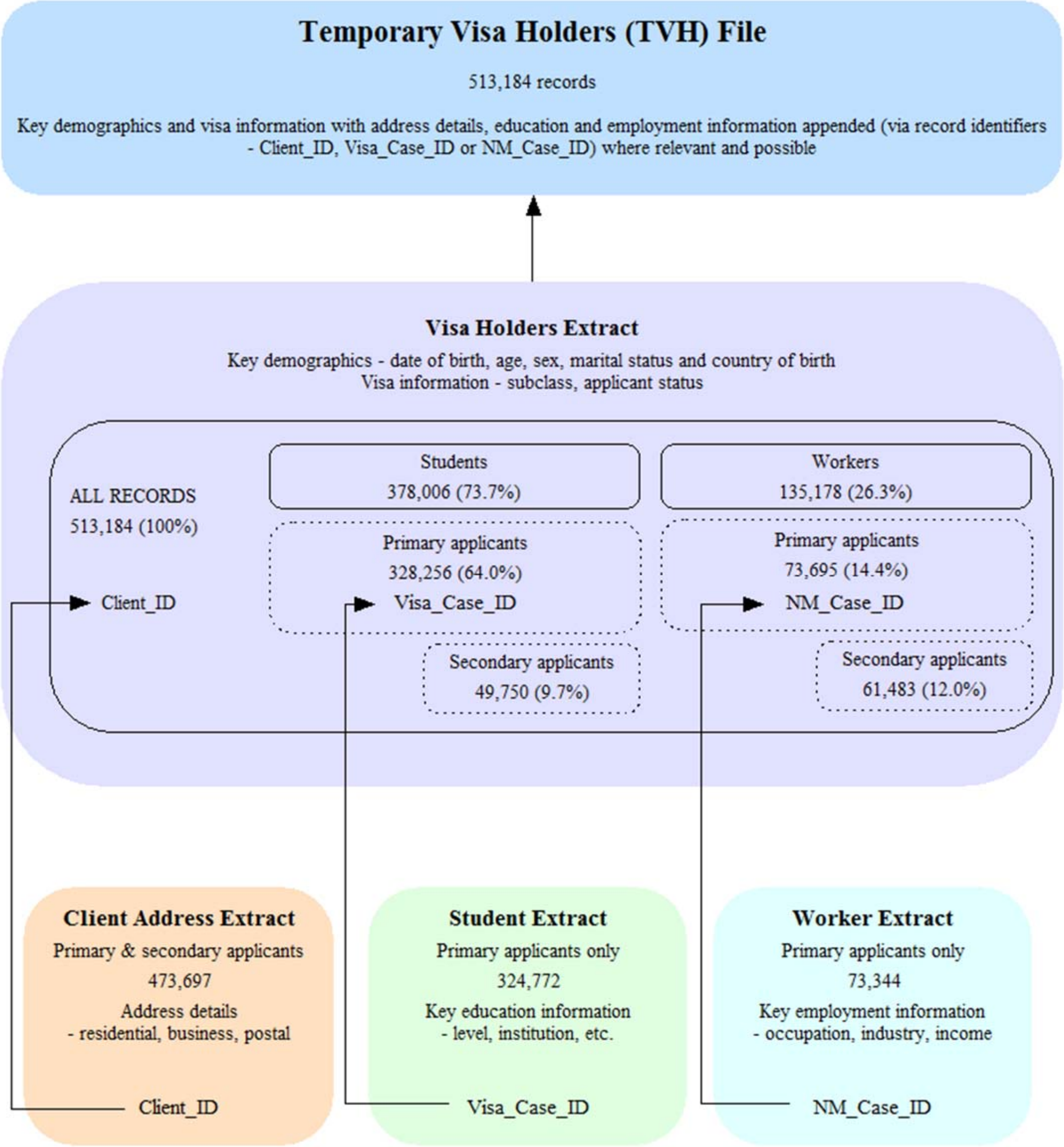
- 570 Independent ELICOS Sector – Full-time English Language Intensive Course for Overseas Students (ELICOS);
- 571 Schools Sector – Full-time Primary or Secondary school course;
- 572 Vocational Education and Training Sector – Main courses of study covered are Certificates I, II, III and IV (except ELICOS), Diplomas, Advanced Diplomas, Graduate Certificates and Graduate Diplomas;
- 573 Higher Education Sector – Main courses of study covered are Bachelor Degrees, Associate Degrees, Graduate Certificates, Graduate Diplomas, Masters by Coursework, Higher Education Diplomas and Higher Education Advanced Diplomas;
- 574 Postgraduate Research Sector – Masters Degree by Research or a Doctoral Degree;
- 575 Non Award Sector – Full-time non-award foundation studies course, or components of a course (other than ELICOS) that does not lead to an award;
- 576 Foreign Affairs or Defence Sector – Allows international students who are sponsored by the Department of Foreign Affairs and Trade or the Department of Defence to study a full-time course of any type.

3.1.4 Temporary work (skilled) visa nominations extract

The Temporary work (skilled) visa nominations extract (the 'Worker extract') contained 73,344 records for those temporary entrants in Australia on a subclass 457 Visa, hereafter referred to as 'Workers'. It pertains to primary applicants only and provides information including the industry and occupation in which the applicant is to be employed, and the remuneration the applicant is expected to receive.

The Subclass 457 Visa allows skilled workers to come to Australia and work for an approved business for up to four years. This visa is designed to enable employers to address labour shortages where they cannot locally source an appropriately skilled Australian resident to do the work.

3.1 Composition of the Temporary Visa Holders dataset


Temporary Visa Holders (TVH) File
513,184 records
Key demographics and visa information with address details, education and employment information appended (via record identifiers - Client_ID, Visa_Case_ID or NM_Case_ID) where relevant and possible
Visa Holders Extract
Key demographics - date of birth, age, sex, marital status and country of birth
Visa information - subclass, applicant status
ALL RECORDS
513,184 (100%)
Client_ID
Students
378,006 (73.7%)
Primary applicants
328,256 (64.0%)
Visa_Case_ID
Secondary applicants
49,750 (9.7%)
Workers
135,178 (26.3%)
Primary applicants
73,695 (14.4%)
NM_Case_ID
Secondary applicants
61,483 (12.0%)
Client Address Extract
Primary & secondary applicants
473,697
Address details
- residential, business, postal
Client_ID
Student Extract
Primary applicants only
324,772
Key education information
- level, institution, etc.
Visa_Case_ID
Worker Extract
Primary applicants only
73,344
Key employment information
- occupation, industry, income
NM_Case_ID

3.1.5 Issues associated with the consolidation of the extracts

Diagram 3.1 shows the relationship between the four component extracts, and resulting consolidated TVH dataset.

Additional information from the Client Address, Student and Worker extracts was appended to the Visa Holders extract through the use of unique record identifiers available on each extract. There were a number of issues encountered in doing this which resulted in additional address, education and/or employment information being unable to be appended for a small number of records.

These were:

- There were 8,016 missing Client IDs and five duplicate Client IDs on the Visa Holders extract. In such cases, an individual's demographic and visa information could not be connected with their address information from the Client Address extract. In the cases where Client IDs were duplicated, it was unclear which of the duplicate records on the Visa Holders extract the information on the Client Address extract related to, so it was safest to disregard this information.
- Similarly, of the 328,256 primary Student records on the Visa Holders extract, there were 26 missing and 17 duplicated Visa Case IDs – meaning no information from the Student extract was able to be appended.
- Finally, of the 73,695 primary 457 applicants on the Visa Holders extract, there were 97 missing and 180 duplicated Nominee Case IDs, which meant that no information from the Workers extract could be appended.

3.2 2011 Census of Population and Housing

The 2011 Census of Population and Housing dataset provides a wealth of information about the Australian community, covering topics such as key demographics, housing, education, participation in the labour force, and occupation and industry of employment. The Census dataset used for this study consisted of 20,928,304 records, from which all name and address information had been removed. Excluded from the dataset were the records of imputed persons and some overseas visitors. Imputed persons are people known to exist, but for whom no Census form was returned. A statistical method is used to impute their demographic information.

From these 20,928,304 Census records, 18,261,808 respondents explicitly indicated they were Australian citizens. Given that temporary entrants do not have Australian citizenship, these records were removed prior to linking. The Census dataset employed in this linkage project comprised the Census records of the remaining 2,666,496 respondents.

4. THE LINKING PROCESS

The statistical linking methodology employed for this project is called probabilistic record linkage (Fellegi & Sunter, 1969). This method links records from two datasets using several variables common to each. A key feature of the 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 compare. This allows ranking of all possible links and optimal assignment of the link or non-link status (Solon and Bishop, 2009).

The probabilistic record linkage methodology used for this project has been broken down into the following steps:

- Standardising the data;
- Choosing blocking variables;
- Selecting linking variables and comparison functions;
- Implementing the record pair comparison; and
- Applying a decision model.

While the same methodological approach has been employed to link all records on the TVH dataset, different information is available for different subpopulations. In particular, additional education and employment information is available for primary applicants, from the Students and Workers extracts (Section 3.1). For this reason, different strategies have been employed to link primary and secondary applicants. Sections 4.1 to 4.5 outline the broad linking methodology, with a focus on the strategies used to link primary applicants. Section 4.6 then summarises the modified strategy used to link secondary applicants.

4.1 Data standardisation

Before records on two datasets are compared, the contents of each need to be made as consistent and comparable as possible. This may involve a number of steps such as verifying, recoding and reformatting data fields, or parsing text fields (i.e. separating text fields into their components).

Some variables differ between the two datasets in a predictable way, and some adjustment is required to remove this difference. Some variables may initially be coded differently on the two datasets, and concordances may be required to align the coding. Variable responses may also be recoded or combined in order to obtain a more robust form of the variable for linking. This set of procedures is termed 'standardisation'.

The standardisation procedures for this project included the coding of imputed and invalid values for all variables to a common missing value to remove the possibility of recording false or spurious agreement on such values.

The following is a description of further standardisation techniques that were performed on selected variables.

Address

Address records on the TVH dataset were converted to the Meshblock (MB), Statistical Area Level 1 (SA1), Statistical Area Level 2 (SA2), Postcode (PCODE) and State levels of the *Australian Statistical Geography Standard* (ASGS) (ABS, 2011) through a geocoding process in order to create geographical linking variables comparable with the Census.

Address information can be critical to establishing reliable record links, and it is highly desirable that any credible source of address information should be exploited. For this reason, all address types (residential, postal and business) except offshore addresses were geocoded and used for linking. Though postal and business addresses (including the addresses of educational institutions) may not be as useful as residential addresses in establishing links, they may still provide a broad indication of where the temporary visa holder is residing, especially in the absence of more specific address information.

Where no address information was available based on the Client Address extract, the 'State of education provider' variable from the Student extract was used to impute the State of residence for primary Student applicants. State was imputed in this way for 42,999 international students (accounting for 13.1% of all primary Student applicants).

Where secondary applicants were missing an address or had an address record that could not be geocoded, the address of the corresponding primary applicant was initially substituted. In practice, it subsequently proved infeasible to link secondary applicants independently, and the final strategy for linking secondary applicants employed the explicit assumption that primary and secondary applicants would reside at the same address (see Section 4.6).

Finally, where persons on the Visa Holders extract of the TVH had multiple address records on the Client Address extract, duplicate records were created containing the differing address information, but replicating all other information associated with the record (date of birth, age, sex, etc.). The same action was applied to records on the Census where the respondent's usual residential address differed from the place of enumeration. This increased the chances that a record would be compared with its true match at some point in the linking process.

Age (AGE)

Date of birth is recorded on the TVH dataset as day, month and year of birth, from which Age on Census night was calculated for use in this study.

Day of birth (BDAY)

Day and month of birth information on both the Census and TVH datasets were combined and recoded to a value of between 1 and 366, corresponding to the 366 unique birthdays within a year.

Marital status (MSTA)

Records on the THV dataset were recoded to align with standard Census categories, and then aggregated into three categories – 'Married', 'Widowed, Divorced and Separated', and 'Never Married'. The marital status of visa holders on the TVH dataset under the age of 15 years on 9 August 2011 was recoded to "Not applicable", to be consistent with the Census coding. It was more problematic to recode the categories of "Engaged" and "*de facto* partner" which are present on the TVH dataset. This is because there is no way to determine whether these people have ever been married, separated, divorced or widowed prior to their engagement or entering into a *de facto* relationship. However, for the purposes of this study these persons were coded to "Never married".

Country of birth (COB1 and COB4)

Responses on the TVH dataset were mapped from four-character texts codes to corresponding four-digit numeric codes in the *Standard Australian Classification of Countries* (SACC) (ABS, 2008b). Country of birth codes at both the one- and four-digit level were utilised in the linking process.

Educational attainment (EDU)

Course level information available on the TVH dataset for Students was used to create the EDU variable. This variable was created to align with the Level of Highest Educational Attainment variable on the Census, broadly describing the highest school or non-school level of education an individual has reported. Records on both the TVH dataset and the Census dataset were then assigned to four broad categories –

1. Year 11 and below,
2. Year 12,
3. Bachelor Degree, Advanced Diploma and Diploma Level, and
4. Postgraduate, Graduate Diploma and Graduate Certificate Level.

Level of educational institution attending (LEV)

Information on course sector, course level and the State/Territory of the course provider, available on the TVH dataset for Students, was combined to create the LEV variable. This variable was coded to align with the Type of Educational Institution

Attending variable on the Census and broadly describes the level/type of education institution an individual was recorded as attending. Records on both the TVH dataset and the Census dataset were then assigned to four broad categories –

1. School,
2. Technical or Further Education Institution,
3. University or Other Tertiary, and
4. Other.

Occupation (OCC)

Roughly 1,000 records on the TVH dataset required concording from the *Australian and New Zealand Standard Classification of Occupations* (ANZSCO) Version 1.2 to *ANZSCO First Edition, Revision 1* (ABS, 2009) as used in the 2011 Census. Resulting codes on both the TVH and Census datasets were then aggregated to the three-digit level for linking purposes.

Industry (IND)

Industry codes on the Census dataset were aggregated to the Division level of the *Australian and New Zealand Standard Industrial Classification (ANZSIC), Revision 1.0* (ABS, 2008a), as this was the finest level available on the TVH dataset.

Income (INC)

Annual remuneration values on the TVH dataset were recoded to align with the standard Census income ranges for Total Personal Income (weekly).

4.2 Blocking variables

Once datasets have been standardised, record pairs (consisting of one record from each dataset) can be compared to see whether they are likely to be a match, i.e. belong to the same person. However, if the datasets are even moderately large, comparing every record on Dataset A with every record on Dataset B is often computationally infeasible, and usually also unnecessary. Blocking reduces the number of comparisons needed by only comparing record pairs where matches are likely to be found – namely, records which already agree on a pre-selected set of blocking variables. Blocking variables are selected based on their reliability and discriminatory power. Sex is only partially useful for example, as it is typically well reported, however it is minimally informative as it only divides the datasets into two blocks, and is thus often used in conjunction with other variables.

Of course, there will be instances where matching records will fail to agree on a selected blocking variable, due to missing, invalid of legitimately different responses on one or the other record. To mitigate this, the linking process is repeated a number of times ('runs'), using a range of different blocking strategies.

Table 4.1 presents the blocking variables used in each run for primary Students and Workers. The strategy for linking secondary applicants is summarised later in Section 4.6.

4.1 Blocking strategy for Primary applicants

	Run							
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Geography	Meshblock	SA1	SA2	PCODE	STATE	STATE	–	–
Birthday	–	–	–	–	BDAY	BDAY	BDAY	–
Birth year	–	–	–	–	BYEAR	BYEAR	BYEAR	–
Age	–	–	–	–	–	–	–	AGE
Sex	–	SEX	SEX	SEX	SEX	–	SEX	SEX
Country of brth 1 digit	–	COB1	COB1	–	–	–	–	–
Country of birth 4 digit	–	–	–	COB4	–	–	–	–
				Students only				
Level of education	–	–	–	–	–	–	–	LEV
				Workers only				
Occupation	–	–	–	OCC	–	–	–	OCC
Industry	–	–	IND	–	–	–	–	IND

To further reduce the computational intensity of linking, the Census dataset was subset before linking primary applicants. For Students, only those records which did not explicitly state that they were not attending an educational institution were retained for use in linking. For Workers, only those Census records which did not explicitly state that they were not in the labour force were retained.

4.3 Linking variables and comparison functions

After the blocking stage has reduced the number of record-pair comparisons down to a computationally feasible level, records from the two datasets are compared using a full suite of linking variables. The choice of linking variables is different for each run and is related to the blocking variables for that run. In addition to standard demographic linking variables, additional education and employment information available on the TVH dataset have been utilised to improve the likelihood of finding unique, high-quality links amongst a population which is characterised by a high degree of homogeneity. This is especially important given that detailed name and address information was not available for use as linking variables.

For each linking variable, a comparison function is used to determine the amount of agreement required between values from the two datasets. The comparison functions used in this study are:

- Exact match (e.g. Sex). Agreement occurs only when the two field values are identical. This criterion is used for most linking fields.
- Numeric difference (e.g. Year of arrival). Two responses may be deemed to agree if their field values differ by an amount less than or equal to a specified maximum difference.

Table 4.2 presents the blocking and linking variables for each run used to link primary applicants, and the comparison functions used for the linking variables. The blocking and linking variables used to link secondary applicants are outlined in Section 4.6.

4.2 Blocking and linking variables for Primary applicants (a)

		Run							
Variable	*Comparison Function*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Geography	Exact match	MB	SA1	SA2	PCODE	STATE	STATE	SA2	PCODE
		–	–	–	–	SA1	SA1	–	–
Birthday	Exact match	BDAY	BDAY	BDAY	BDAY	BDAY	BDAY	BDAY	BDAY
Birth year	Exact match	–	–	–	–	BYEAR	BYEAR	BYEAR	–
Age	Exact match	AGE	–	AGE	AGE	–	–	–	AGE
Age	Num. diff. (±1 year)	–	AGE	–	–	–	–	–	–
Sex	Exact match	–	SEX	SEX	SEX	SEX	SEX	SEX	SEX
Marital status	Exact match	MSTA	MSTA	–	MSTA	MSTA	MSTA	MSTA	MSTA
Country of birth (1)	Exact match	–	COB1	COB1	–	–	–	–	–
Country of birth (4)	Exact match	COB4	COB4	COB4	COB4	COB4	COB4	COB4	COB4
Year of arrival	Num. diff. (±1 year)	YOA	YOA	YOA	YOA	YOA	YOA	YOA	YOA
		Students only							
Education	Num. diff. (±1)	EDU	EDU	EDU	EDU	EDU	EDU	EDU	EDU
Level of education	Exact match	LEV	LEV	LEV	LEV	LEV	LEV	LEV	LEV
		Workers only							
Occupation	Exact match	–	–	OCC	OCC	–	OCC	OCC	OCC
Industry	Exact match	IND		IND	IND	IND	–	–	IND
Income	Num. diff. (±1)	INC	–	INC	INC	INC	INC	INC	INC

(a) Shading denotes blocking variables.

The different runs feature a trade-off between the level of agreement on geographical variables (Meshblock, SA1, etc.) versus personal characteristics (such as Birthday, Country of Birth, etc.) required to establish credible links. They also make use of additional employment and education information for the relevant subpopulations on the TVH dataset.

The numeric difference comparison function has been used in some runs, for particular variables, to account for legitimate differences in the reporting of these variables between the two datasets as outlined below.

Year of arrival (YOA)

The Census Year of Arrival variable was compared with the closest equivalent on the TVH dataset – the date an individual's visa was granted. As such, there are conceptual differences between these two variables. The Year of Arrival variable on the Census is the year a person first arrived in Australia to live for one year or more. There are a number of reasons why this may not match up precisely with the date a visa was granted on the TVH dataset. For example, a period of time may elapse between a visa being granted and a temporary visa holder arriving in Australia, or a temporary entrant may have first arrived in Australia at an earlier date on a different visa.

Reflecting these sources of potential difference, the numeric difference comparison function was utilised to allow a visa grant date within ± one year of the Census Year of Arrival variable to constitute a match.

Educational attainment (EDU)

The Educational Attainment variable on the TVH dataset reflects the level of education that a person is currently undertaking, while on the Census, it reflects the level of education a person has already attained. As such, the numeric difference comparison function has been utilised to account for potential legitimate differences between the two datasets. For example, this would allow a record on the Census with an attained level of education of Year 12 to be considered a link with a record on the TVH dataset which reports currently undertaking a Bachelor Degree.

Income (INC)

The income figure derived from the TVH dataset reflects the gross earnings an employer expects to pay the primary 457 visa holder, including any fringe benefits. This figure is based on the time the nomination was approved, which may be up to four years prior to the 2011 Census. There are a number of legitimate reasons that this variable may differ to the income variable on the Census. For example, the 457 visa holder may be receiving a higher rate of remuneration at the time of the Census due to salary indexing.

For these reasons, the numeric difference comparison function was applied to the income variable, allowing an income range within ± one to constitute a match.

4.4 Record pair comparison

Table 4.2 details, for each proposed probabilistic record linkage run, the selected blocking and linking variables and the comparison functions to be employed for each linking variable.

The blocking variable(s) define the *comparison space* of the run – that is, the list of all *record pairs* (one record from each dataset) that will be considered. Conceptually, this comparison space can be sorted into two disjoint sets – the set of matching pairs (in which both records relate to the same individual) and the set of non-matching pairs.

For every record pair within the comparison space it is possible to construct a *field comparison vector* that reports whether or not the paired records agree or disagree on each of the selected linking variables, using the associated comparison functions as the basis for determining agreement.

While knowledge of the field comparison vectors can quickly identify the most likely matches (records that agree on all linking variables) and non-matches (records that disagree on all linking variables), it does not provide an adequate basis for ranking intermediate outcomes.

For this, it is necessary to establish measures of the significance of agreement on each linking variable relative to all other linking variables. These measures – *field weights* – must reflect the probability that a variable comparison for matching records may be distorted by error or non-response and the probability that a non-matching record pair may agree by chance on a given variable.

Record pair weights, formed by combining the *field weights* of all linking variables, then provide a measure of the joint likelihood of realising the observed field comparison vectors. Specifically high record pair weights correspond to patterns of concordance that are extremely unlikely to be observed for record pairs that are not in fact a true match. The record pair weights thus provide a basis for ranking all record pairs within the comparison space from most probable matches to most probable non-matches.

Underlying the computation of field weights (and record pair weights) are the theoretical concepts of *m-probabilities* and *u-probabilities*:

$$m = P\left(\text{field agrees} \middle| \text{records belong to the same individual}\right)$$

$$u = P\left(\text{field agrees} \middle| \text{records belong to different individuals}\right)$$

Neither of these probabilities can be computed empirically prior to the conduct of a probabilistic linking run – as their definitions imply knowledge of the set of matching record pairs. (This knowledge may not even be available following the completion of the data linking run.) Hence, implementation of the Fellegi-Sunter model relies upon the use of adequate *estimates* of the necessary m- and u-probabilities.

It is frequently feasible to compute accurate estimates of the required u-probabilities based upon the entire comparison space – since the removal of the unidentified set of matching pairs would result in no significant difference to the calculation. Unfortunately, no equivalent simple strategy exists for the computation of reliable m-probabilities.

In previous studies, the ABS has used an in-house implementation of the expectation-maximisation (EM) algorithm (Samuels, 2012) to estimate m- and u-probabilities. For this study, a more empirical estimation approach was deemed necessary, as several important features of the datasets required closer scrutiny.

Whereas the earlier feasibility study (ABS, 2014) implicitly assumed that the majority of primary applicants on the TVH dataset would have responded to the Census, and that the address information included in the TVH dataset would provide a reliable guide for linking, initial data investigations challenged these assumptions. The comparison spaces defined by the chosen blocking variables clearly could not deliver the expected numbers of matching record pairs.

Estimation of the m- and u-probabilities was achieved by means of thorough data confrontation within the comparison spaces for all linking runs, reviewing carefully the likely numbers of matching record pairs and the reliability of key linking variables (especially address information).

It has been observed that u-probabilities in particular can vary significantly with respect to selected subpopulations within the two datasets. By exploiting this information, and independently linking selected subpopulations, it is possible to extract superior information with which to classify probable matches and non-matches. Table 4.3 defines the broad population characteristics that were used for this purpose in linking primary applicants.

4.3 Demographic categories used to define significant subpopulations for primary applicants

Age cohort	*Country of birth*	*Region*
A1=Aged under 25	C1=Born in the United Kingdom	R1=Does not live in a major city
A2=Aged over 25	C2=Born in China	R2=Lives in a major city
	C3=Born in India	
	C4=Born elsewhere overseas	

While the focus of m- and u-probabilities is generally upon the probability of agreement (and disagreement) on a given variable within matching and non-matching record pairs, it can also arise that a variable comparison returns a result of 'missing'. This occurs where either or both of the paired records has a missing response to the variable in question. Missingness is generally assumed to be uninformative in probabilistic linking, implying that the m- and u-probabilities of missingness are roughly equal, but it is still important to consider and account for these probabilities in the estimation process.

Following the Fellegi-Sunter linking methodology, m- and u-probabilities may be converted to field weights via the following formulae:

$$\text{Agreement weight} = \log_2\left(\frac{m}{u}\right);$$

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

$$\text{Disagreement weight} = \log_2\left(\frac{1-m-m_{\text{missing}}}{1-u-u_{\text{missing}}}\right).$$

From the above equations, it is easy to deduce several intuitively desirable properties of the Fellegi–Sunter framework:

- Agreement weights are always positive and disagreement weights are always negative;
- Missing weights should be approximately, if not exactly zero;
- The magnitude of the agreement weight is driven primarily by the likelihood of chance agreement (i.e. a low probability of chance agreement results in a high agreement weight);
- And conversely, the magnitude of the disagreement weight is driven primarily by the stability and reliability of the linking variable (i.e. disagreement on a variable that is consistently and accurately reported will be heavily penalised).

Information on the blocking and linking variables, the choices of agreement comparison functions and the estimated m- and u-probabilities are entered into linking software. The ABS uses a modified version of the open source data linking software Febrl (Christen, Churches and Hegland, 2004) to carry out probabilistic record linkage.

4.5 Applying a decision model

Once all candidate record pairs have been generated, a decision rule is required to determine which record pairs, or links, should be retained in the output dataset. In this study, a two-stage process was employed. In the first stage, which was largely automated, only those record pairs that satisfied an optimal one-to-one assignment process were retained. In the second stage, which is more subjective, a decision was made regarding which of the remaining links were most credible and fit-for-purpose.

4.5.1 One-to-one assignment

Initially, all linking runs were conducted independently. The subsequent process of identifying optimal one-to-one links proceeded as follows:

1. The links generated by each run were assigned adjusted aggregate link weights to permit comparability between runs, and all such links were then collated.

2. Where two records were linked within multiple runs, only the instance with the highest adjusted link weight was retained.

3. For each record in the TVH dataset, a search was conducted for a unique best link to a Census record. This required that *both* linked records were identified as the unique best link to the other. (At this stage, primary applicant links that were supported by credible evidence from the secondary applicant linking process were always preferred – see Section 4.6.). All unique best links were then extracted, and all remaining links involving the extracted records were deleted.

4. The optimal assignment of one-to-one links among the remaining record pairs could not be achieved through simple inspection, as complex many-to-many relationships were present. Instead, a linear optimisation algorithm (Christen and Churches, 2005) was employed to provide a solution.

4.5.2 Cut-off weights

The second phase of the decision rule stage takes the output of the 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-off weights.

The cut-offs for this project were determined by clerically reviewing record pairs to identify where cut-offs weights were best positioned to trade off the probability of missing true matches against the probability of including false links.

A clerical reviewer is often able to utilise information, which cannot be captured in the automated comparison process, such as common transcription errors (e.g. 1 and 7, 5 and 6). However, clerical review is a resource intensive task, and given the limited resources available for this quality study, targeted clerical review was undertaken. Record pairs were ordered by adjusted aggregate weight and a random sample was selected for review based upon the particular weights which were contributing a large number of links. More attention was targeted around the lower end of the weight distribution, with more record pairs inspected to fine-tune the cut-offs.

As a result of this process, two potential cut-off weights (17 and 23) for primary applicant links were identified for further review. Links with a weight of 23 and above were assessed as suitable for inclusion in the output dataset without further inspection. Similarly, links with a weight lower than 17 were deemed to have insufficient proof of match status – and most were demonstrably incorrect.

Closer examination of links with weights between 17 and 20 revealed that additional rules could be devised to determine which links should be retained and which should be rejected. Within this weight range, Student links were only retained if a link was found for an associated secondary applicant, or if the record pair agreed on specific linking variables. All links with a weight between 20 and 23 were retained despite some reservations about their reliability.

The varying quality of links should be noted and considered by potential users. It is almost certain that no single dataset will be optimal for all forms of analysis. Some applications may require a highly accurate linked dataset, with very few false links, whereas others will benefit from a dataset which is more representative of the temporary entrant population. In such cases, the inclusion of representative false links may be less damaging to analytical goals than the omission of some difficult-to-link demographics. Further discussion of the quality of links is contained in Section 5.

4.6 Summary of the linking process for Secondary applicants

The information available on the TVH for secondary applicants is restricted to basic demographic variables and geographic information of varying quality (as discussed in Sections 3.1.2 and 4.1). As such, a separate linking strategy was developed for secondary applicants that utilised the strength of the links generated for associated primary applicants, for which additional education and employment variables were available to assist in identifying record pairs.

This strategy utilised two 'Case ID' variables on the TVH dataset, which are shared between members of the same family or migratory unit. For Students, primary and associated secondary applicants may share the same 'Visa Case ID'. It should be noted however, that there are some exceptions to this rule. All applicants holding Student visas that are part of the same family or migratory unit share a common Visa

Case ID only when all applications are made at the same time. Where a primary Student applicant arrives first, and family members subsequently apply for a visa to reunite with the primary applicant, they have a different Visa Case ID. Workers from the same family or migratory unit always share the same 'Nominee Case ID'.

From the three best quality links generated for each primary applicant (those above the cut-off of 17 rules as discussed in Section 4.5.2), an index was created between their Case ID (either Visa or Nominee) and the Dwelling IDs on the Census records to which potential links were found. These Dwelling IDs were then appended to all secondary applicant records that shared the same Case ID, and were used as a blocking variable. This meant that links for secondary applicants were only searched for within the Census dwelling/s in which the associated primary applicant had potentially been found.

The blocking and linking strategy for secondary applicants is presented in table 4.4.

4.4 Blocking and linking variables for Secondary applicants (a)

Variable	*Comparison function*	
"Geography"	Exact match	Dwelling ID
Birthday	Exact match	BDAY
Age	Numeric difference (±1 year)	AGE
Sex	Exact match	SEX
Country of birth (4 digit)	Exact match	COB4
Year of arrival	Numeric difference (±1 year)	YOA

(a) Shading denotes blocking variables.

The methodology and strategies for secondary applicants then followed those outlined in Sections 4.4 and 4.5. A single cut-off was defined for both Students and Workers.

Where links for secondary applicants of sufficient quality (i.e. above the cut-off) were found, this information was used to determine which link was retained for the associated primary applicant. All primary links for which an associated secondary applicant was found within the same Dwelling ID were assigned a higher aggregate link weight before the primary links underwent the process of identifying optimal links (outlined in Section 4.5.1) to ensure the appropriate links were selected and retained on the final linked dataset.

Once all primary and secondary links had been processed, the properties and quality of the linked dataset could be assessed. Quality assessment includes an examination of the unlinked records and an assessment of the amount of under- and over-representation of subgroups on the linked dataset.

5. EVALUATION OF THE LINKAGE

At the completion of the linkage process, 243,991 (48%) of the 513,184 records from the TVH dataset were linked to a 2011 Census record to create the ACTEID. The linkage rate achieved varied for different subpopulations on the TVH dataset. For example, close to 70% of Workers on the TVH were linked to a Census record, compared with 41% for Students. A summary of linkage rates for key subpopulations is provided in table 5.1.

5.1 ACTEID linkage rates, by subpopulation

Visa type / Applicant type	*TVH Dataset* *(no.)*	*ACTEID* *(no.)*	*Linkage rate* *(%)*
International students			
Primary	328,256	135,079	41.2
Secondary	49,750	20,788	41.8
Total	378,006	155,867	41.2
457 workers			
Primary	73,695	48,329	65.6
Secondary	61,483	39,795	64.7
Total	135,178	88,124	65.2
Total temporary entrants			
Primary	401,951	183,408	45.6
Secondary	111,233	60,583	54.5
Total	513,184	243,991	47.5

Table 5.2 summarises the quality of the links generated for primary applicants, based on the following two factors:

1. Whether or not a secondary applicant associated with the primary applicant (i.e. sharing the same Visa or Nominee Case ID) was linked. Those primary links that are supported by secondary links are considered to be of the highest quality, given the extra evidence that a secondary applicant being found in the same dwelling provides.
2. For those links which were not supported by secondary applicant links, the weight range (as discussed in Section 4.5.2) that they fell within, which reflects the level of agreement between linking variables on the two datasets.

5.2 Quality of primary applicant links

	Students		*Workers*	
	(no)	*(%)*	*(no.)*	*(%)*
Associated secondary applicants linked	14,681	10.9	20,468	42.4
Weight of 23 or above	71,001	52.6	18,050	37.3
Weight of 20 to 23	30,194	22.4	4,356	9.0
Weight of 17 to 20	19,203	14.2	5,455	11.3
Total	135,079	100.0	48,329	100.0

The distribution of links among the quality categories defined in table 5.2 differs for the Student and Worker subpopulations. The proportion of primary worker applicants for which associated secondary applicants were linked is markedly higher than for Students. This largely reflects the different demographics of these two cohorts, with Students generally being younger than Workers, and less likely to be married. As such, Students are more likely to temporarily migrate to Australia on their own, while Workers are more likely to bring their families. This is evident when looking at the ratio of primary to secondary applicants. Only one in every 11.5 primary Student applicants were accompanied by secondary applicants, compared with one in every 2.4 primary Worker applicants.

Irrespective of these differences, the majority of both Student and Worker links fell into the top two quality categories, having an associated secondary applicant/s linked or having a weight of 23 or above, indicating a high level of agreement between key variables on the TVH and Census datasets.

The quality of the linked dataset was further evaluated on a number of measures:

- The number of links against the expected number of links,
- The characteristics of linked records,
- The properties of the TVH records that did not get linked to a Census record, and
- The under- or over-representation of subgroups in the linked dataset.

5.1 Comparing expected number of links to actual number of links

An important aspect of understanding the linkage results is considering how many TVH records we might reasonably expect to link to the Census. Persons on the TVH dataset may be missing from the Census for a number of reasons, including:

- They did not realise they should have completed the Census form;
- The temporary entrant was not in education or employment at Census time;
- Detailed information was not collected because they were residing in a Non-Private Dwelling (e.g. hotel or student accommodation);
- They were temporarily out of the country on Census night (out of scope of the Census);
- They had not yet arrived in Australia, despite being granted a visa; or
- They may have been unable to complete a Census form, particularly when their first language is not English.

The previously discussed feasibility study (see Section 2) noted that the linkage rate attained for this project could be considerably lower than predicted (70%) if certain subpopulations of temporary entrants failed to respond to the Census, irrespective of their eligibility. For the feasibility study, it was implicitly assumed that the majority of primary applicants on the TVH dataset would have responded to the Census, however the findings from this study challenge that assumption, suggesting that many of the visa holders on the TVH dataset were simply not on the Census to be found.

Results from the Census Post Enumeration Survey (PES) show that certain groups of people are more likely to be missed by the Census based on characteristics such as their country of birth, age, sex and geographic location. For more information, see *Census of Population and Housing – Details of Undercount, 2011* (ABS, 2012).

Countries of birth found to have particularly high rates of undercount in the 2011 Census included China, India and the Philippines. Young adults, particularly males from 20 to 24 and 25 to 29 years of age, were also generally harder to capture in the Census, irrespective of their residency status or country of birth. These subpopulations make up large proportions of the full TVH dataset as shown in table 5.3, and these factors likely significantly affected the linking outcomes, particularly for the international student cohort.

5.3 All TVH records, by selected countries of birth and age groups

	International students		*457 workers*	
	(no.)	*(%)*	*(no.)*	*(%)*
		Country of Birth		
China	92,566	24.5	4,922	3.6
India	50,454	13.3	16,911	12.5
Philippines	5,248	1.4	10,748	8.0
		Age group		
Under 15 years	17,249	4.6	25,505	18.9
15–19 years	41,684	11.0	3,846	2.8
20–24 years	165,011	43.7	5,042	3.7
25–29 years	94,399	25.0	27,387	20.3
30–34 years	37,828	10.0	28,275	20.9
35+ years	21,835	5.8	45,123	33.4

In addition to being missed by the Census, some temporary entrants may have responded but been identified as 'Overseas Visitors' based on their response to the question "Where does the person usually live?". For the 2011 Census, overseas visitors were those people who indicated they would be usually resident in Australia for less than a year. The only variables available for 'Overseas Visitors' in Census output are Age, Sex, Marital Status and geographic information based on place of enumeration. As such, there may not have been enough descriptive variables to establish good quality links for these records. There were 189,145 records on the Census subset used for this study identified as overseas visitors, accounting for 7.1% of the subset.

5.2 Characteristics of linked records

Table 5.4 shows the proportion of linked primary applicant records that agreed on various linking variables, where a response was not missing on either dataset.

5.4 Proportion of linked primary applicant records agreeing on linking variables (a)

	Students		*Workers*	
	(no)	*(%)*	*(no.)*	*(%)*
Meshblock	65,046	56.3	21,481	54.4
Statistical Area Level 1	69,084	59.6	22,597	56.9
Statistical Area Level 2	89,006	76.7	27,506	69.2
Postcode	93,515	77.1	28,737	69.6
State	130,548	97.5	39,449	95.5
Birthday	129,527	100.0	44,849	98.8
Birth year	131,532	97.4	46,498	96.3
Age	133,608	99.0	47,462	98.3
Age ±1 year	134,074	99.3	47,909	99.2
Sex	132,744	99.6	47,817	99.5
Marital status	124,792	95.4	43,105	91.8
Country of birth (1 digit)	125,256	99.1	46,341	99.1
Country of birth (4 digit)	124,501	98.5	46,013	98.4
Year of arrival ±1 year	81,647	68.0	33,798	74.5
Educational attainment ±1	106,918	96.0	–	–
Level of education undertaken	99,124	82.5	–	–
Industry	–	–	26,175	56.6
Occupation	–	–	30,786	66.1
Income±1	–	–	41,661	90.0

(a) Calculated as a proportion of total linked records minus those that had missing information.

The lowest rates of agreement were generally found for geographic variables, however at least 95% of both Student and Worker links agreed on State. This is indicative of the issues surrounding address information on the TVH discussed earlier, and may also suggest high levels of mobility amongst temporary entrants in Australia.

Very high levels of agreement (>96%) were found for the core demographic variables of Birthday, Birth Year, Age, Sex and Country of Birth. Lower rates were achieved for variables where there may have been legitimate changes/differences between the TVH and Census datasets, such as Marital Status and Year of Arrival.

Relatively high rates of agreement were found for the Education variables however, the high level of aggregation of these variables limited their ability to help distinguish links amongst the highly homogenous Student population.

Employment information was found to be very useful in distinguishing links amongst the Worker population, due to the higher level of granularity available for these variables (i.e. roughly 100 occupations and 20 industries). However, relatively lower rates of agreement were achieved for the Occupation (66%) and Industry (57%) variables when compared with key demographic variables. Clerical review of selected record pairs revealed that while these variables might not have been an exact match at the three-digit ANZSCO occupation level or Division level of the ANZSIC, there was often still some similarity in the information provided. For example, there were instances of occupation being listed as 851 Food Preparation Assistants on one file and 351 Food Trades Workers on the other, or 224 Information and Organisation Professionals on one file and 551 Accounting Clerks and Bookkeepers on the other, etc. This may be reflective of the different ways in which information on the TVH and Census is collected and/or coded. For example, the TVH form may be filled out by a third party i.e. the sponsor of the 457 visa applicant or a Migration Agent, while the Census is generally filled out by the visa holder themselves, or a member of their household. Differences in the way these individuals may interpret and respond to questions on the TVH and Census datasets may help to explain the relatively lower rates of agreement seen for the Occupation and Industry variables.

5.3 Characteristics of unlinked TVH records

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

- The corresponding Census record does not exist (as discussed in Section 5.1),
- There is insufficient data on either the TVH or corresponding Census record to adequately link (i.e. missing information), or
- The quality of data on either the TVH or corresponding Census record is too poor to facilitate linking.

Data quality can be affected by respondents not supplying all of the information requested, errors in the information provided or coding errors. These issues affect both the TVH and Census datasets, and may be exacerbated by low levels of English language proficiency as mentioned previously, or transcription errors. Table 5.5 compares the number of records with missing information for the 218,543 primary TVH records which were not linked to a Census record to the 138,408 that were linked.

Table 5.5 shows that there are generally low levels of missing responses on the TVH dataset for key blocking and linking variables, with the exception of geographic variables. As discussed in Sections 3.1.2 and 4.1, addresses on the TVH dataset were geocoded to standard ABS statistical geography to enable comparison with the Census.

5.5 Missing variable frequencies on TVH dataset, primary applicants

	Unlinked TVH records		*Linked TVH records*	
	(no.)	*(%)*	*(no.)*	*(%)*
Meshblock	48,173	22.0	28,466	15.5
Statistical Area Level 1	47,487	21.7	27,700	15.1
Statistical Area Level 2	47,487	21.7	27,700	15.1
Postcode	40,889	18.7	20,755	11.3
State	10,347	4.7	8,169	4.5
Birthday	19	0.0	6	0.0
Birth Year	1	0.0	0	0.0
Age	1	0.0	0	0.0
Sex	0	0.0	0	0.0
Marital Status (age ≥ 15 years)	3,157	1.5 (a)	3,347	1.8 (b)
Country of Birth	375	0.2	395	0.2
Year of Arrival	0	0.0	0	0.0

(a) Based on 217,685 unlinked primary applicant records of persons aged 15 years and over.

(b) Based on 182,117 linked primary applicant records of persons aged 15 years and over.

The quality of the address information on the TVH dataset varied, resulting in varying levels of success in geocoding. Of all unlinked records on the TVH dataset, at least one geographic variable (MB, SA1, SA2, PCODE or STATE) was missing from 48,173, or 22%, of records. This is a higher rate than that observed for linked primary applicant records (roughly 15%). This demonstrates the impact of having good quality and complete address information on the linking outcome.

A number of options to deal with missing/inadequate address information were discussed in the aforementioned feasibility study (see Section 2), such as conducting the linkage only on those records that were able to be geocoded to a Meshblock. For completeness, the linkage outlined in this paper was conducted on the full TVH dataset, with all available address information and additional education and employment linking variables utilised to increase the likelihood of finding good quality links (in the absence of detailed address information).

Another notable feature of the unlinked TVH records is the higher proportion of primary applicants without associated secondary applicants. For Students, 93.9% of unlinked primary Student applicants did not have any associated secondary applicants compared with 87.6% of linked primary Students, while for Workers; the difference was even more pronounced (70.7% of unlinked records compared with 52.4% of linked records).

5.4 Under- and over-representation of subgroups

The realised link rate of (243,991 / 513,184 =) 48% was not uniformly achieved across all subpopulations of interest within the TVH dataset. When comparing the composition of the ACTEID with the full TVH dataset, the analysis revealed that the linking process missed significant numbers of records for some subpopulations. In particular, lower than average link rates were found for temporary entrants born in China, those aged 18 to 25 years, those not married and those granted a visa closer to Census day. Many of these findings align with known population characteristics of Census undercount. For more detail about 2011 Census undercount, see *Census of Population and Housing – Details of Undercount, 2011* (ABS, 2012). Additionally, the following Student and Worker subpopulations of note were found to be under-represented:

Students

- People born in India,
- People obtaining Year 12 educations, and
- Those attending technical or further education (TAFE) institutions.

Workers

- People born in the United States of America,
- People employed in the Arts and Recreation Services, Other Services and Information Media and Telecommunications industries, and
- People with medium and low skilled occupations.

See tables A.1 and A.2 in the Appendix for more information on the link rates realised for selected subpopulations.

Section 6 of this Paper explains how a weighting strategy can be used to mitigate the proportional under- and over-representation of demographic subpopulations on the linked dataset and improve the utility of the data for statistical analyses.

6. WEIGHTING STRATEGY

To ensure that the linked ACTEID dataset is appropriate for statistical analysis, adjustments need to be made to account for the under- and over-representation of demographic subpopulations as described in Section 5. It is recommended that this is achieved through the use of a weighting strategy. This is a well-established method of accounting for variations in representation, which is a feature common to all data linkage projects.

It should be noted however that the use of a weighting strategy to restore representativeness presupposes that there are always persons in the linked dataset who are sufficiently similar to those persons who have not been linked. This cannot be guaranteed for linked data, since there may be specific characteristics that result in certain individuals not being linked. In addition, it may be prudent to question the wisdom of weighting unconvincing links.

For this study, a two stage weighting strategy for primary applicants was employed to restore the representativeness of the linked ACTEID dataset. Different weighting strategies will suit different analytical purposes, however the strategy outlined below can provide a guide for more refined weighting strategies to enable more detailed analysis of subpopulations of interest.

The crucial first stage weights identify and adjust for any relative under- or over-representation of key subpopulations within the linked dataset. The second stage weights were applied to fine-tune or calibrate selected weighted aggregates to corresponding totals from the complete dataset for analytical variables of interest.

6.1 Demographic categories used to define subpopulations for stage 1 weights

		Students	
Sex	*Age Cohort*	*Marital Status*	*Level of Education*

S1=Male
S2=Female
	A1=Under 25 years
	A2=25 years & over
		M1=Married
		M2=Not married
			L1=Infants / Primary & Secondary school
			L2=Technical & Further Education
			L3=University & Other Tertiary
			L4=Other
			L5=Missing

		Workers	
Sex	*Age Cohort*	*Marital Status*	*Country of Birth*

S1=Male
S2=Female
	A1=Under 30 years
	A2=30 to 39 years
	A3=40 years & over
		M1=Married
		M2=Not married
			R1=United Kingdom(a)
			R2=Rest of Europe
			R3=India
			R4=Rest of Asia
			R5=Sub-Saharan Africa
			R6=Other(b)

(a) Includes Channel Islands and Isle of Man.

(b) Includes North Africa and the Middle East, the Americas, Oceania and Antarctica, missing and inadequately described etc. responses.

6.1 Stage 1 weights

The design of the Stage 1 weighting strategy is based on the categories shown in table 6.1.

Selecting the categories shown in table 6.1 and deciding how to partition them required a degree of judgement. Too many categories with too many partitions can generate very small subpopulations for which the results may prove spurious in the sense that consistent results might not be observed if the linkage were conducted on another dataset of TVH records. Conversely, too few categories can fail to adequately capture the diversity of achieved linking outcomes. The Stage 1 weighting strategy for this study was developed through a process of informed and methodical examination, with similar strategies aimed for in weighting both the Students and Workers.

During the Stage 1 weighting process, both Students and Workers were initially aligned to Sex by Age cohort by Marital Status totals from the true (original) TVH dataset. However, given the differences in under- and over- representation between the Student and Worker populations, an additional, independent weighting category was also employed to further restore representativeness for key population attributes. Thorough investigations found that Level of Education for Students and Country of Birth for Workers produced the best results for this purpose.

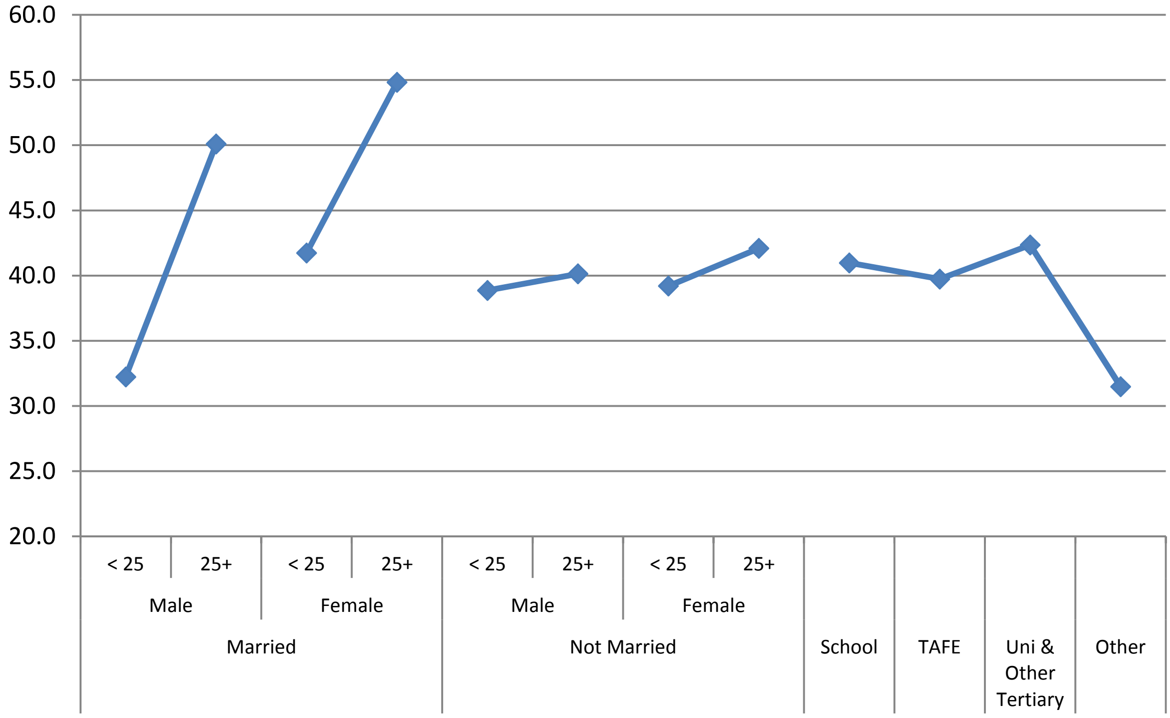
After applying Stage 1 weights, the linked dataset for Students aggregated to the (2 × 2 × 2 =) 8 Sex by Age cohort by Marital status totals and the five Level of Education totals from the TVH dataset. For Workers, the linked dataset aggregated to the (2 × 3 × 2 =) 12 Sex by Age cohort by Marital status totals and the six Country of Birth totals.

Figures 6.2 and 6.3 report the observed linked rates from the linked dataset for the (8 + 5 =) 13 Student and (12 + 6 =) 18 Worker reference subpopulations mentioned above.

Significant differences can be observed in the link rates for the subpopulations displayed in figures 6.2 and 6.3, which both highlights some systematic patterns in the link rate outcomes and confirms the suitability of the Stage 1 weighting categories. Looking at Students for example, the TVH records for females, aged over 25 years, married and attending a University or Other Tertiary education institution have generally linked more successfully than other demographic groups. For Workers, females, aged 30 to 39 years, married and from the United Kingdom and Sub-Saharan Africa (largely driven by South Africa) have achieved higher link rates.

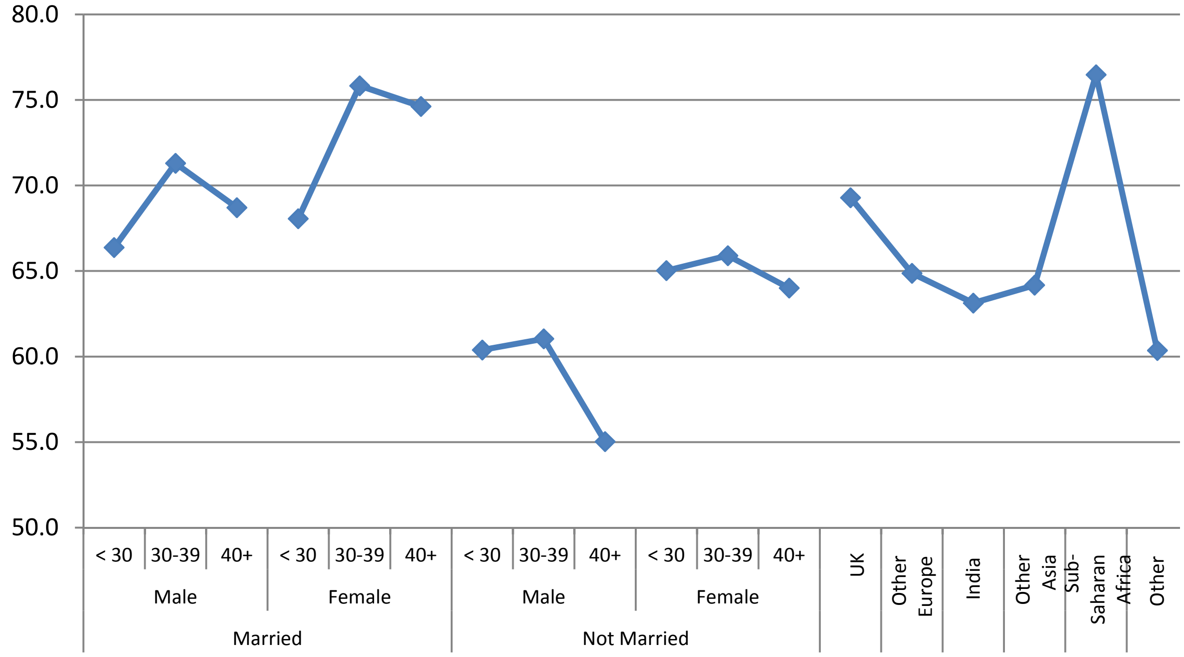
The Stage 1 weights have been derived from the reciprocals of these link rates. Thus, subpopulations with lower link rates are assigned higher weights, increasing their representation within the weighted linked dataset.

6.2 Link rates for 13 Student subpopulations (a)



(a) Link rates for missing Level of Education not displayed.

6.3 Link rates for 18 Worker subpopulations



A simplified example of weighting follows. The TVH dataset contains 13,594 records for married males aged 30 to 39 years in Australia on a 457 visa. The linked dataset contains 9,692 of these individuals. Therefore, the first stage weight required to restore the representativeness of this demographic subpopulation is (13,594 / 9,692=) 1.40.

The Stage 1 weights applied to Student records ranged from a low of 1.65 (for married females aged 25 years and over and attending a University or Other Tertiary institution) to a high of 3.35 (for unmarried females aged under 25 years and attending an ‘Other’ educational institution). For Workers, weights ranged from 1.12 (for married females aged from 30 to 39 years from Sub-Saharan Africa) to 1.97 (for unmarried males aged 40 years and over from ‘Other’ countries of birth).

6.2 Calibration of Stage 1 weights

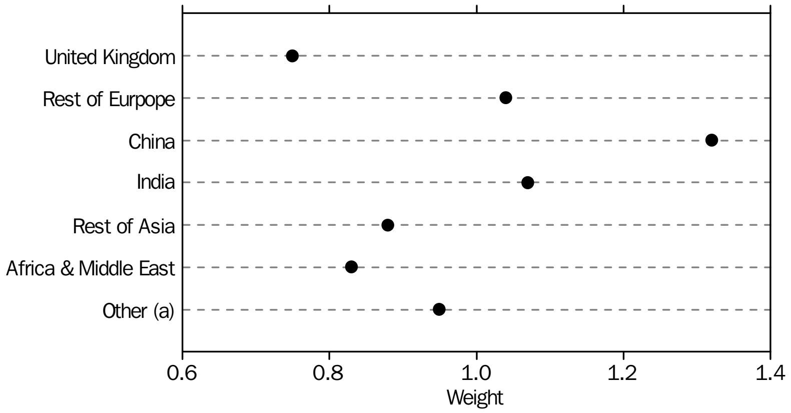
Stage 1 weights may be modified to derive Stage 2 weights where there is a need to present weighted aggregates that correspond exactly to the record counts in selected categories in the original TVH dataset, or where there are some question regarding the ability of the Stage 1 weights to restore sufficient representativeness for important analytical dimensions of the linked dataset. Analysis of weighted data after the application of Stage 1 weights indicated that there was still an element of under- or over-representation for some key analytical variables, and therefore there was a need for Stage 2 weights.

Figures 6.4 and 6.5 illustrate these considerations. Figure 6.4 highlights the remaining variation in weighted Country of Birth totals for Students after applying Stage 1 weights. It shows the Stage 2 weights required to align the weighted Country of Birth totals from the linked dataset to the Country of Birth totals on the complete TVH dataset.

While Stage 1 weights were able to reproduce Country of Birth totals for some categories (those close to 1.0 in figure 6.4), several categories require considerably higher or lower weights to be applied.

Students who were born in China require the highest Stage 2 weights while those born in the United Kingdom require the lowest. This large variation may be explainable by differences in the demographic, geographic and educational characteristics associated with specific countries of birth, and varying proportions of difficult to link subpopulations.

6.4 Students – Stage 2 weights required to align with Country of Birth benchmarks



(a) Includes the Americas, Oceania and Antarctica, inadequately described and missing responses.

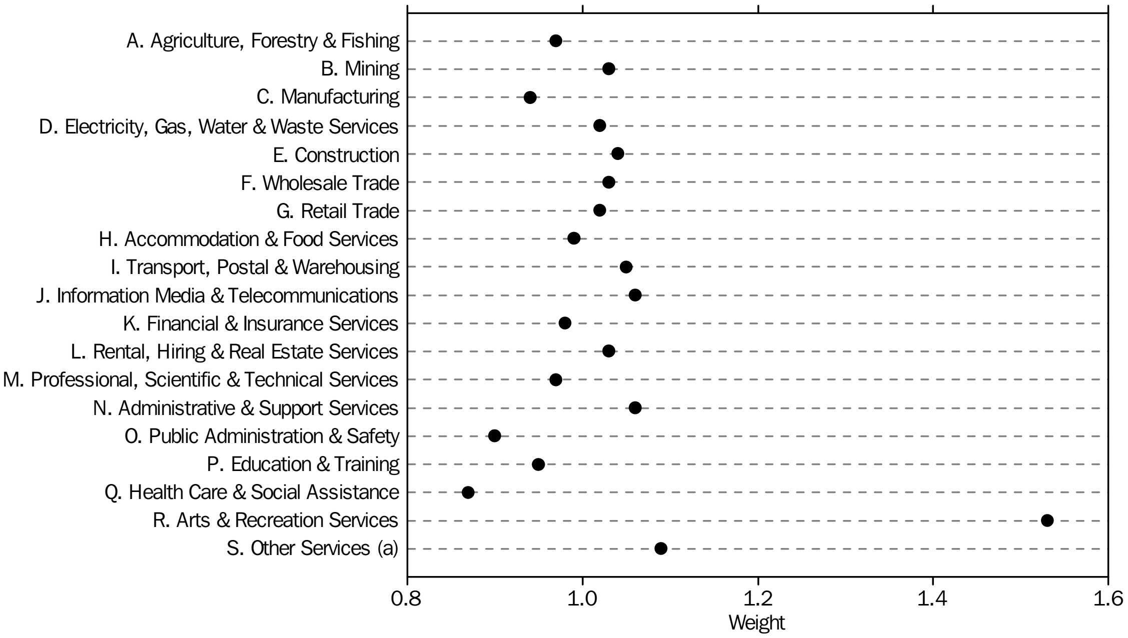
China had the lowest link rate of any Country of Birth for primary Student applicants, and is noted as a country of birth that is particularly hard to enumerate in the Census. Students from China on the TVH were also largely concentrated in the Under 25 years of age group and had a higher than average proportion who were not married. These subpopulations were also found to be amongst the most difficult to link. A combination of these factors has most likely contributed to the low link rate for China, resulting in higher than average weights being required.

In contrast, Students from the United Kingdom attained the highest link rate. Most were aged over 25 years and were attending University or Other Tertiary institutions. Students attending University or Other Tertiary institutions and aged over 25 years of age had higher link rates than those attending other education institutions and aged under 25 years. There was also a higher than average proportion of Students from the United Kingdom residing in Western Australia, which had a comparatively higher link rate than the other States and Territories.

Figure 6.5 illustrates the Stage 2 weights required to calibrate the weighted Industry division counts derived from the linked dataset for Workers with those on the complete TVH dataset.

It can be observed that Stage 1 weights have reproduced the Industry totals for most categories reasonably well, since the corresponding Stage 2 weights are clustered around one. However, some variation is observable and it should be noted that the sample sizes associated with some categories are quite small. This implies that some variation should be anticipated. The Arts and Recreation Services industry requires the highest Stage 2 weight, while the Health Care and Social Services industry requires the lowest. As with the previous example, differences in the demographic characteristics (e.g. age and sex) associated with specific industries, and varying proportions of difficult to link subpopulations explains much of this variation.

6.5 Workers – Stage 2 weights required to align with Industry benchmarks



(a) Includes missing responses.

Compared with all primary Workers, those in the Arts and Recreation Services industry had a higher than average proportion in younger age groups, who had never been married and who had a year of arrival closer to the Census date. These groups were all found to be particularly difficult to link, hence achieving below average linkage rates. The distribution of those in the Health Care and Social Assistance industry across core demographic variables was much more consistent with the average for all primary Workers, with one exception. The proportion of females in the Health Care and Social Assistance was more than double that seen on the full TVH file. A greater proportion of those in this industry also came from countries of birth that achieved linkage rates at the higher end of the scale.

6.3 Implementation of Stage 2 weights

When designing the Stage 2 weighting strategy, consideration was given to additional variables likely to be of analytical interest. The variables decided upon are shown in table 6.6.

After the application of Stage 2 weights, the linked dataset for Students aggregated to the (2 × 6 × 7 =) 84 Sex by Age cohort by Country of Birth totals. For Workers, the linked dataset aggregated to the (2 × 3 × 7 =) 42 Sex by Age cohort by Occupation Group totals, with an additional, independent adjustment made to aggregate to the 19 Industry totals from the TVH dataset.

6.6 Demographic categories used to define subpopulations for Stage 2 weights

Students

Sex	*Age Cohort*	*Country of Birth*
S1=Male	A1=20 years & under	R1=United Kingdom (a)
S2=Female	A2=21 to 22 years	R2=Rest of Europe
	A3=23 to 24 years	R3=China
	A4=25 to 26 years	R4=India
	A5=27 to 28 years	R5=Rest of Asia
	A6=29 to 30 years	R6=Africa & the Middle East
	A7=30 years & over	R7=Other (b)

Workers

Sex	*Age Cohort*	*Occupation Group*	*Industry Division (c)*	
S1=Male	A1=Under 30 years	O1=Managers	I1=A	I11=K
S2=Female	A2=30 to 39 years	O2=Health Professionals	I2=B	I12=L
	A3=40 years & over	O3=Design, Engineering, Science & Transport Professionals	I3=C	I13=M
			I4=D	I14=N
		O4= Other Professionals	I5=E	I15=O
		O5=Technicians & Trades Workers	I6=F	I16=P
		O6=Medium & Low Skilled Occupations	I7=G	I17=Q
		O7=Missing	I8=H	I18=R
			I9=I	I19=S(d)
			I10=J	

(a) Includes Channel Islands and Isle of Man.

(b) Includes the Americas, Oceania and Antarctica, missing and inadequately described etc. responses.

(c) See figure 6.5 for Industry Division titles.

(d) Includes missing responses.

When applying Stage 2 weights, counts for the 13 Student and 18 Worker subpopulations as identified for Stage 1 no longer aligned with the totals in the TVH dataset as they would if only Stage 1 weights were applied.

Table 6.7 illustrates the impact of second stage weighting for a small subset of the data – Male workers aged 30–39 years employed in Manager occupations. It can be seen that the second stage weights restore the linked dataset for all Industries of employment reported on the TVH dataset (i.e. 3,598). This was only approximately true for the first stage weights.

Note that the totals for individual industries are not exactly aligned by the second stage weighting for the selected subpopulation shown in table 6.7. However, they will typically be better aligned than under Stage 1 weighting, and are exactly aligned when summed across all 42 Sex by Age cohort by Occupation Group subpopulations.

6.7 Weighted estimates, by selected Industries – Males aged 30 to 39 years, Managers

	Unweighted record counts		*Weighted estimates*	
Industry	*TVH Dataset*	*ACTEID*	*Stage 1*	*Stage 2*
Information Media & Telecommunications	381	221	326	356
Financial & Insurance Services	311	218	315	315
Accommodation & Food Services	307	192	300	300
Construction	274	181	258	284
Mining	258	168	243	265
...	...	...	...	...
All Industries	3,598	2,322	3,414	3,598

The desirability of calibrating to Industry benchmarks will vary depending on the type of analysis users intend to perform i.e. whether the focus is on a specific Industry or whether it involves comparisons between multiple Industries. If the analytical focus is on the characteristics within a specific industry of employment, scaling to Industry totals may potentially distort analysis. This highlights the careful consideration potential users will need to give in deciding which weighted estimates to use.

There are a number of possible cross-tabulations that could be performed with the new ACTEID dataset. While the two-stage weighting strategy described in this paper is designed to be a good, general purpose approach, it does not adjust for all possible analytical dimensions in the data.

7. APPLICATIONS / POTENTIAL ANALYSES

While the purpose of this paper is to describe the linking methodology and report on the quality of the resulting linked ACTEID dataset, it is worth briefly emphasising the utility of the linked dataset in enabling a better understanding of temporary entrant populations.

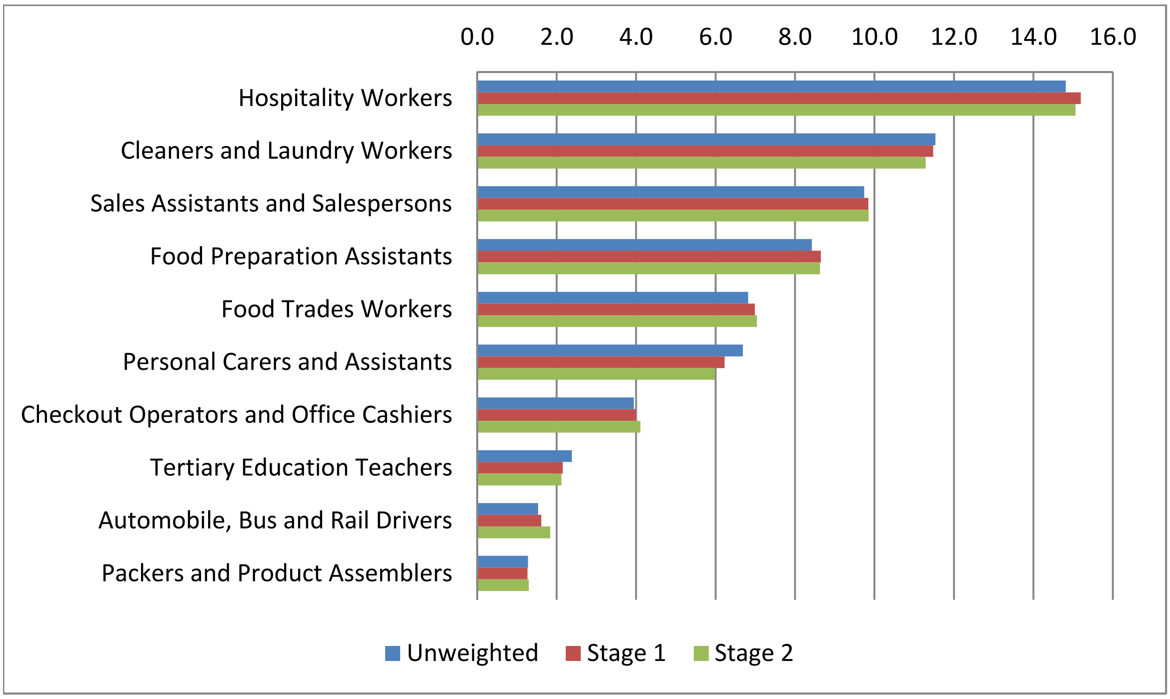
ACTEID provides information on the socio-economic and geographic characteristics of international students and 457 visa holders in much more detail that other currently available datasets. Current sources of information include:

- Administrative data on the basic characteristics and stocks of international students and 457 visa holders. Examples include:
  - Information collected by the Department of Education and Training on enrolments and commencements of international students (Department of Education and Training, 2016), and
  - Quarterly reports on the number of 457 visas granted published by DIBP (DIBP, 2016b).
- Relatively small sample surveys and cases studies. Examples include:
  - The Department of Education and Training's International Students Survey with a sample size of roughly 56,000 students in 2014 (Department of Education and Training, 2015), and
  - The DIBP's 2012 Survey of Subclass 457 Employers and Employees with a sample size of roughly 3,800 457 visa holders and 1,600 457 visa employers (DIBP, 2014).

ACTEID offers a broader range of data items than other datasets, and is larger (i.e. over 240,000 records) enabling analysis at a finer level of detail than possible from sample surveys. In this respect, ACTEID not only provides a useful dataset in its own right but can also complement other existing administrative data relating to international students and 457 visa holders held by other Government departments. ACTEID will allow policy makers and researchers to delve much deeper into the social and economic characteristics of the growing temporary entrant population in Australia than previously possible.

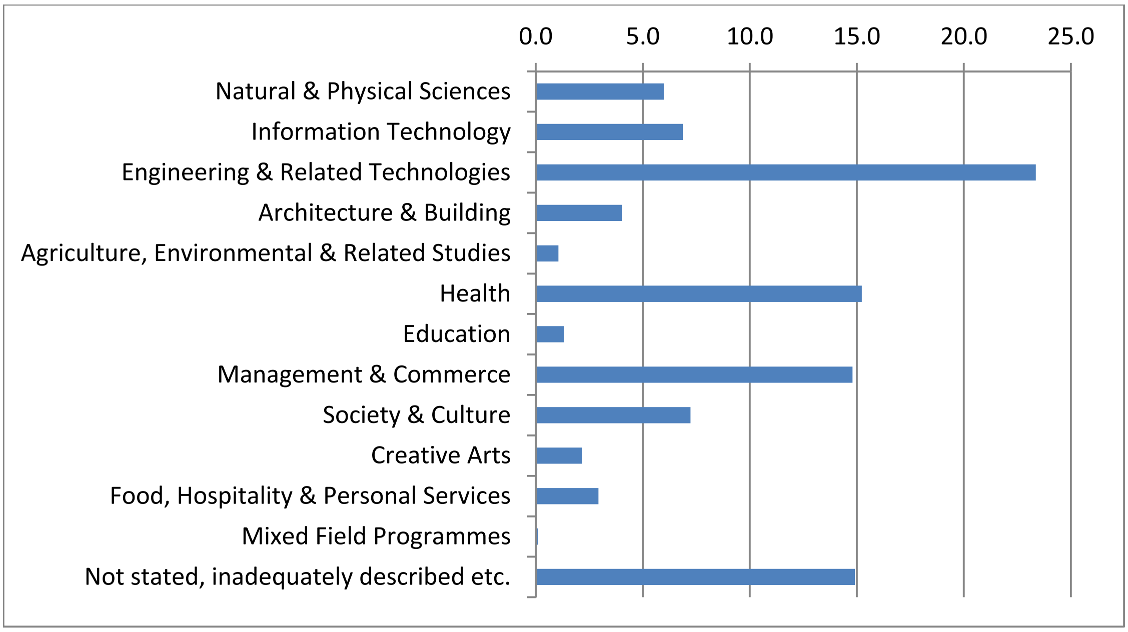
The example below provides a brief insight into the economic and labour market contributions made by international students while studying in Australia. Just over a third of all linked primary applicant Students (46,685, or 34.6%) identified themselves as being employed in the labour force, based on unweighted data. The top 10 occupations of employment reported are shown in figure 7.1, with both unweighted and weighted estimates presented to provide an indication of the effects of weighting on analyses. The most commonly reported occupations were Hospitality Workers, Cleaners and Laundry workers, and Sales Assistants and Salespersons.

7.1 Proportion of employed primary applicant international students in top 10 occupations



The new linked dataset also enables investigations into the qualifications held by 457 Workers. To demonstrate this, Figure 7.2 shows unweighted estimates for the broad field of study of the highest completed non-school qualification held by primary 457 applicants.

7.2 Proportion of primary applicant 457 Workers, by broad field of education of highest completed non-school qualifications



For those 457 workers holding qualifications in the three most common fields of study, the most frequently reported occupations (unweighted) are displayed in table 7.3. This provides some insight into whether 457 Workers are employed in jobs commensurate with their skills.

7.3 Most commonly reported occupations of primary 457 Workers, by field of education

Occupation	*(no.)*	*(%)*
Field of Education: Engineering and Related Technologies		
Engineering Professionals	3,010	26.9
Business and Systems Analysts, and Programmers	894	8.0
Construction, Distribution and Production Managers	804	7.2
Total	11,208	100.0
Field of Education: Health		
Medical Practitioners	3,536	48.2
Midwifery and Nursing Professionals	2,389	32.6
Health Therapy Professionals	404	5.5
Total	7,331	100.0
Field of Education: Management and Commerce		
Accountants, Auditors and Company Secretaries	1,167	16.5
Sales, Marketing and Public Relations Professionals	548	7.7
Business Administration Managers	505	7.1
Total	7,083	100.0

Further considerations for potential analyses are beyond the scope of this paper.

8. RECOMMENDATIONS FOR FUTURE LINKAGES

The main aim of this study was to gain an in-depth understanding of the quality of the integrated dataset resulting from the linkage between the TVH and Census datasets, and any issues that may arise from under- and over-representation of certain subpopulations. It has been shown that a valuable dataset which supports appropriate statistical analysis can be produced and as such, plans are underway to refine the experimental dataset to allow further analytical investigation. Given the promising results achieved through this study, it is recommended that the linkage be repeated utilising the 2016 Census and a 2016 TVH extract, and it is anticipated that appropriately confidentialised data from this linked dataset be made available publically to enable users to conduct analysis and research. This section describes a number of recommendations which should be considered in conducting any future linkages.

As discussed in Section 3.1.2, a number of issues were identified with the Client Address extract of the TVH dataset and a number of actions were taken to mitigate them. All address types were geocoded, and where address information for Students was missing or incomplete, information about the state / territory of an education institution was substituted. Where clients recorded multiple addresses, multiple records were created. A similar procedure was applied to the Census where individuals had a different place of usual residence and place of enumeration. The aim of creating multiple records with differing address information was to improve the likelihood of a record being paired with its true match in the linking process.

To align non-geographic (demographic and population specific) information, careful data standardisation and repair was undertaken on both datasets. In addition, population specific information such as education information for Students and employment information for Workers was used to increase the number of available linking fields. However, these interventions provide only a partial mitigation to respondents supplying inaccurate or incomplete information. Factors that might affect quality of reported information may include English proficiency barriers, proxies completing visa applications or Census forms on behalf of the visa holder, and scanning/ imputation errors with Census data.

The extensive efforts outlined above to maximise the utility of information on both datasets underscores the importance of having linking fields that contain unique and uncommon values available in probabilistic linking. In this respect, there is likely to be significant benefit to the quality of future Census to TVH linkages (and therefore the quality of decisions made using the resultant statistics) from the use of anonymised name and address information in the linkage. The retention of name and address information from the 2016 Census for up to four years may facilitate this.

Being able to use anonymised name and address information will provide a greater range of linking fields that contain more discriminating information for individual records. Students comprise the majority of records on the TVH dataset and mostly report a marital status of ‘never married’. The possibility of having detailed name information would provide more relatively unique and stable information about an individual (that would unlikely change in the time between the extraction of TVH data and the Census enumeration period) and could have the potential to improve the quality of the statistics produced from the dataset by both improving the quality of linked record pairs, and the number of records linked.

With these considerations in mind, it is recommended that future TVH to Census linkage projects as a minimum should:

1. Adopt a subpopulation linking strategy as outlined in this paper, to account for the availability of different information as it relates to Students and Workers;
2. Consider the use of anonymised name and address information, as linking fields;
3. Adopt an address repair strategy as outlined in this study to fully utilise all address information for geocoding;
4. Continue to explore potential to improve the quality of address information on the TVH i.e. requiring regular address updates; and
5. Continue to develop strategies to emphasise and reinforce the requirement of temporary entrants to participate in the Census and provide accurate information.

9. CONCLUSION

Following on from the Feasibility Study in 2014, the objective of this study was to link the TVH dataset to the 2011 Census, and assess the quality of the new linked dataset to provide a new and previously unavailable source of information for research and analytical purposes. In addition to the quality issues surrounding address information identified in the previous feasibility study, the linkage exercise proved particularly challenging as a large proportion of the TVH pertains to components of the population that are characterised by a large degree of homogeneity and are especially difficult to enumerate.

To overcome these challenges, the linking strategy devised for the feasibility study was reviewed and expanded upon to increase the likelihood of finding high quality, unique links. Key features of the expanded strategy were the addition of education and employment information as linking variables and the use of a dwelling indicator to assist in linking secondary applicants.

The resulting linked dataset contains over 240,000 records out of a possible 513,000, equating to an overall linkage rate of 48% (41% for Students, and 65% for Workers). The main reasons that remaining records were unable to be linked were that a corresponding Census record did not exist or that the quality of information on the TVH and/or Census was not of sufficient quality to enable linking. Nevertheless, ACTEID provides users with a rare opportunity to uncover and understand the socio-economic characteristics of temporary entrants in Australia. In this regard, ACTEID is able to provide a detailed insight into these populations in a way that has not previously been possible. However, it is critically important that users understand the nature and limitations of this probabilistically-linked dataset.

The use of cut-offs as described in Section 4.5 involved an informed trade-off between the ‘representativeness’ and ‘precision’ of the final linked dataset. Importantly, users must recognise that ACTEID is not perfectly representative of the entire temporary visa holder population. Of particular note, persons born in China, young adults aged roughly 18 to 25 years and those that were not married were found to be especially under-represented on the linked dataset. These subpopulations make up large proportions of the TVH dataset, and particularly the international student cohort. Section 6 discusses the use of weights which are required to restore representativeness among key subpopulations.

The addition of migration characteristics from the TVH to the social and economic information from the Census is a significant development in the mission to improve and expand on the range of migrant statistics. Information about the increasing temporary entrant population is a known data gap. This dataset shows much promise in being able to provide new insights into how temporary entrants engage with the Australian labour market, their impact on addressing skill shortages and their economic contributions. In this regard, ACTEID can enable the development and evaluation of more informed migration policy. Future enhancements to improve ACTEID data quality could be gained through the expanded use of name and address information.

REFERENCES

Australian Bureau of Statistics (2008a) *Australian and New Zealand Standard Industrial Classification (ANZSIC), Revision 1.0*, cat. no. 1292.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1292.0 >

—— (2008b) *Standard Australian Classification of Countries (SACC)*, cat. no. 1269.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1269.0 >

—— (2009) *ANZSCO – Australian and New Zealand Standard Classification of Occupations, First Edition, Revision 1*, cat. no. 1220.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1220.0 >

—— (2011) *Australian Statistical Geography Standard (ASGS): Volume 1 – Main Structure and Greater Capital City Statistical Areas*, cat. no. 1270.0.55.001, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.001 >

—— (2012) *Census of Population and Housing – Details of Undercount, 2011*, cat. no. 2940.0, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/2940.0 >

—— (2014) "Assessing the Suitability of Temporary Migrants Administrative Data for Data Integration", *Methodology Research Papers*, cat. no. 1351.0.55.053, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.053 >

Christen, P.; Churches, T. and Hegland, M. (2004) "Febrl – A Parallel Open Source Data Linkage System", *Proceedings of the Eighth Pacific–Asia Conference, PAKDD 2004*, Sydney, Australia, pp. 638–647.

Christen, P. and Churches, T. (2005) *Febrl 0.3 Documentation.* < http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/ >

Department of Education and Training (2015) *International Student Survey 2014 Overview Report, April 2015.* < https://internationaleducation.gov.au/research/research-papers/Documents/ISS%202014%20Report%20Final.pdf >

—— (2016) *International Student Data.* < https://internationaleducation.gov.au/research/international-student-data/pages/default.aspx >

Department on Immigration and Border Protection (2014) *Filling the Gaps: Findings from the 2012 Survey of Subclass 457 Employers and Employees*. < https://www.border.gov.au/ReportsandPublications/Documents/research/filling-gaps.pdf >

—— (2016a) *Temporary Entrants and New Zealand Citizens Report as at 30 June 2016.* <https://www.border.gov.au/ReportsandPublications/Documents/statistics/br0169-30-june-2016.pdf>

—— (2016b) *Subclass 457 Quarterly Reports,* < http://www.border.gov.au/about/reports-publications/research-statistics/statistics/work-in-australia >

Fellegi, I.P. and Sunter, A.B. (1969) "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64(328), pp. 1183–1210.

National Statistical Service (2010) "*High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes*", February 2010. <http://www.nss.gov.au/nss/home.NSF/pages/High+Level+Principles+for+Data+Integration+-+Content?OpenDocument >

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, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1352.0.55.120 >

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, ABS, Canberra. < http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.025 >

< All URLs last viewed on 23 January 2017 >

APPENDIX

A. LINK RATES FOR SELECTED SUBPOPULATIONS

A.1 Link rates, selected Student subpopulations

	TVH Dataset *(no.)*	*ACTEID* *(no.)*	*Linkage rate* *(%)*
Applicant status			
Primary	328,256	135,079	41.2
Secondary	49,750	20,788	41.8
All records	378,006	155,867	41.2
Primary applicants			
Sex			
Male	174,562	70,415	40.3
Female	153,694	64,664	42.1
Age			
0–9 years	357	221	61.9
10–14 years	1,753	1,070	61.0
15–19 years	40,662	15,968	39.3
20–24 years	160,630	61,988	38.6
25–29 years	81,170	34,737	42.8
30–34 years	28,183	13,242	47.7
35–39 years	9,819	4,880	49.7
40 years and over	5,682	2,973	52.3
Marital status			
Married	52,115	25,677	49.3
Not married	269,720	106,052	39.3
Not stated	6,421	3,350	52.2
Region of birth			
Oceania & Antarctica	2,152	1,178	54.7
North-West Europe	12,688	5,310	41.9
Southern & Eastern Europe	4,518	1,815	40.2
North Africa & the Middle East	14,897	7,535	50.6
South-East Asia	70,492	33,435	47.4
North-East Asia	129,227	44,945	34.8
Southern & Central Asia	63,630	27,166	42.7
Americas	20,980	9,020	43.0
Sub-Saharan Africa	7,351	4,141	56.3
Not stated	2,321	534	23.1
Top 10 countries of birth			
China (excludes SAARs & Taiwan)	90,435	27,659	30.6
India	38,794	15,305	39.5
Malaysia	17,443	8,230	47.2
Korea, Republic of (South)	16,828	7,674	45.6
Vietnam	16,453	6,934	42.1
Indonesia	11,782	6,305	53.5
Thailand	11,006	4,458	40.5
Hong Kong (SAR of China)	10,982	4,865	44.3
Nepal	10,085	4,910	48.7
Singapore	7,451	3,913	52.5

A.1 Link rates, selected Student subpopulations – continued

	TVH Dataset (no.)	*ACTEID (no.)*	*Linkage rate (%)*
Primary applicants			
Year of arrival			
2007 & prior	5,710	2,670	46.8
2008	30,873	14,326	46.4
2009	75,692	31,478	41.6
2010	108,081	44,552	41.2
2011	107,900	42,053	39.0
Education type			
Year 11 & below	1,225	562	45.9
Year 12	16,165	6,195	38.3
Bachelor Degree, Adv. Diploma, Diploma & Certificates	217,764	91,477	42.0
Postgraduate, Graduate Diploma & Graduate certificate	65,245	27,976	42.9
Not stated	27,857	8,869	31.8
Level of education			
Infants/Primary & Secondary school	16,287	6,674	41.0
Technical or Further Education (TAFE)	86,656	34,432	39.7
University or other tertiary	206,277	87,354	42.3
Other	11,615	3,657	31.5
Not stated	7,421	2,962	39.9
State			
New South Wales	115,335	44,680	38.7
Victoria	98,544	41,716	42.3
Queensland	50,913	21,032	41.3
South Australia	19,664	8,326	42.3
Western Australia	28,389	13,122	46.2
Tasmania	2,983	1,424	47.7
Northern Territory	1,255	489	39.0
Australian Capital Territory	7,633	3,155	41.3
Other Territories	2	0	0.0
Not stated	3,538	1,135	32.1

A.2 Link rates, selected Worker subpopulations

	TVH Dataset	*ACTEID*	*Linkage rate*
	(no.)	*(no.)*	*(%)*
Applicant status			
Primary	73,695	48,329	65.6
Secondary	61,483	39,795	64.7
All records	135,178	88,124	65.2
Primary applicants			
Sex			
Male	54,783	35,557	64.9
Female	18,912	12,772	67.5
Age			
0–24 years	2,770	1,582	57.1
25–29 years	20,404	13,000	63.7
30–34 years	20,232	13,490	66.7
35–39 years	12,550	8,560	68.2
40–44 years	8,140	5,633	69.2
45–49 years	4,571	3,004	65.7
50 years and over	5,028	3,060	60.9
Marital status			
Married	32,954	23,177	70.3
Not married	38,509	23,864	62.0
Not stated	2,232	1,288	57.7
Region of birth			
Oceania & Antarctica	613	387	63.1
North-West Europe	28971	19718	68.1
Southern & Eastern Europe	2346	1374	58.6
North Africa & the Middle East	1237	797	64.4
South-East Asia	9642	6348	65.8
North-East Asia	5906	3532	59.8
Southern & Central Asia	11354	7281	64.1
Americas	8387	5047	60.2
Sub-Saharan Africa	4562	3489	76.5
Not stated	677	356	52.6
Top 10 countries of birth			
United Kingdom, Channel Islands & Isle of Man	17,591	12,188	69.3
India	9,848	6,217	63.1
Philippines	5,870	3,919	66.8
Ireland	5,546	3,860	69.6
United States of America	4,359	2,447	56.1
South Africa	3,072	2,424	78.9
China (excludes SAARs & Taiwan)	2,808	1,684	60.0
Germany	2,020	1,280	63.4
Canada	1,923	1,226	63.8
France	1,705	1,120	65.7
Year of arrival			
2007	2,176	1,504	69.1
2008	12,142	8,708	71.7
2009	12,043	8,546	71.0
2010	24,784	16,517	66.6
2011	22,550	13,054	57.9

A.2 Link rates, selected Worker subpopulations – continued

	TVH Dataset	*ACTEID*	*Linkage rate*
	(no.)	*(no.)*	*(%)*
Primary applicants			
Occupation group			
Managers	11,530	7,443	64.6
Health professionals	9,101	6,995	76.9
Design, engineering, science & transport professionals	9,587	6,593	68.8
Other professionals	22,905	14,501	63.3
Technicians & trades workers	15,350	9,748	63.5
Medium & low skilled occupations	4,637	2,743	59.2
Not stated	585	306	52.3
Industry			
Agriculture, Forestry & Fishing	1,599	1,069	66.9
Mining	5,379	3,466	64.4
Manufacturing	5,228	3,639	69.6
Electricity, Gas, Water & Waste Services	1,317	862	65.5
Construction	7,712	4,844	62.8
Wholesale Trade	1,408	906	64.3
Retail Trade	2,576	1,681	65.3
Accommodation & Food Services	3,872	2,501	64.6
Transport, Postal & Warehousing	1,159	734	63.3
Information Media & Telecommunications	7,303	4,459	61.1
Financial & Insurance Services	3,979	2,706	68.0
Rental, Hiring & Real Estate Services	1,396	876	62.8
Professional, Scientific & Technical Services	5,992	4,034	67.3
Administrative & Support Services	481	304	63.2
Public Administration & Safety	888	654	73.6
Education & Training	3,282	2,246	68.4
Health Care & Social Assistance	9,981	7,577	75.9
Arts & Recreation Services	1,009	419	41.5
Other Services	7,910	4,783	60.5
Not stated	1,224	569	46.5
State			
New South Wales	20,823	14,278	68.6
Victoria	12,318	8,776	71.2
Queensland	10,081	7,142	70.8
South Australia	1,963	1,464	74.6
Western Australia	11,479	8,214	71.6
Tasmania	362	261	72.1
Northern Territory	814	546	67.1
Australian Capital Territory	871	609	69.9
Other Territories	6	5	83.3
Not stated	14,978	7,034	47.0

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

© Commonwealth of Australia 2017

Produced by the Australian Bureau of Statistics