1351.0.55.060

Australian Bureau of Statistics

Research Paper

Personal Income Tax and Migrants Integrated Dataset (PITMID) 2011-12 Quality Assessment

New Issue

Research Paper

Personal Income Tax and Migrants Integrated Dataset (PITMID) 2011-12 Quality Assessment

National Migrant Statistics Unit

Population & Social Statistics Division

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 19 OCT 2016

ABS Catalogue no. 1351.0.55.060

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.

PERSONAL INCOME TAX AND MIGRANTS INTEGRATED DATASET (PITMID) 2011–12 QUALITY ASSESSMENT

National Migrant Statistics Unit
Australian Bureau of Statistics

EXECUTIVE SUMMARY

The Australian Bureau of Statistics (ABS) created the Personal Income Tax and Migrants Integrated Dataset (PITMID) by linking the Australian Taxation Office (ATO) Personal Income Tax (PIT) records with migrant records from the Australian Government's Settlement Database (SDB). The PITMID Project initially began in 2013 with a linking feasibility study. During the study, almost a million migrant settlement records (54%) linked to a PIT record demonstrating that the linking was feasible. The study concluded that the linked 2009–10 and 2010–11 PITMID dataset provides valuable new information on recent permanent and provisional migrant taxpayers' personal income.[1] In 2015, the 2009–10 and 2010–11 PITMID data was released in *Personal Income of Migrants, Australia, Experimental* (ABS cat. no. 3418.0).

PITMID contains key personal income variables (employee income, own unincorporated business income, investment income, other income and foreign income) and SDB variables (visa subclass, application status (primary or secondary), location (onshore or offshore), country of birth and year of arrival for Skill, Family, Humanitarian, Other permanent and provisional visa holders). The SDB records are linked to the PIT records using variables such as name, date of birth and address. Relevant legislation and guidelines, including the *Privacy Act 1988* and the *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes* were adhered to, protecting the privacy of individuals on both datasets.

This PITMID study was conducted to assess the effects of the change in the linking methodology introduced in 2016 for the 2011–12 PITMID linkage. The 2009–10 and 2010–11 PITMID linkage employed a combined deterministic and probabilistic linking methodology. The new linking methodology utilises a Statistical Analysis Software (SAS) macro known as the Deterministic linking Macro (D-MAC) for a purely deterministic approach. The D-MAC links two datasets using a simple set of rules and then outputs linked record pairs with a calculated measure of accuracy. The study briefly outlines the original and new linking methodologies and presents the results of the analyses conducted to assess the quality of the 2011–12 PITMID linkage compared

1 For further information, see *Feasibility Study of Linking Migrant Settlement Records to Personal income Tax Data* (ABS cat. no. 1351.0.55.051) and *Personal Income of Migrants, Australia, Experimental* (ABS cat. no. 3418.0)

with the 2009–10 and 2010–11 PITMID linkages. This was done by running the D-MAC over the full SDB dataset and the 2009–10 and 2010–11 PIT datasets.

The new methodology utilising the D-MAC was found to be much quicker to administer and produced high quality results, while enabling comparison between the annual series. The linking results generated by D-MAC showed almost 95% of the SDB records either linked to the same PIT record (as the previous linking) or did not link to a PIT record. For this reason, the links generated for 2009–10 and 2010–11 were retained for the 2011–12 linkage process. It is anticipated that the PITMID Project will continue to use the D-MAC for linking in future. The D-MAC is also becoming the preferred linking method for other important ABS data integration projects. Utilisation of the same linking methodology for PITMID will ensure that the project is well placed should any further opportunities arise for linking with other datasets in the future.

ACKNOWLEDGEMENTS

This paper was prepared by the ABS' National Migrants Statistics Unit (NMSU) with assistance from the Data Linkage Centre (DLC) and the Data Integration, Access and Confidentiality Methodology Unit (DIACMU). The NMSU work program is jointly funded by the Australian Bureau of Statistics, the Department of Immigration and Border Protection (DIBP) and the Department of Social Services (DSS).

The ABS acknowledges the assistance provided to the PITMID Project by the Department of Immigration and Border Protection (DIBP), the Department of Social Services (DSS) and the Australian Taxation Office (ATO).

The results of this study are based, in part, on tax data supplied by the ATO to the ABS under the *Taxation Administration Act 1953*, which requires that such data is only used for the purpose of administering the *Census and Statistics Act 1905*. Any discussion of data limitations or weaknesses is in the context of using the data for statistical integration purposes, and is not related to the ability of the data to support the ATO's core operational requirements.

Legislative requirements to ensure privacy and secrecy of this data have been adhered to. In accordance with the *Census and Statistics Act 1905*, results have been confidentialised to ensure that they are not likely to enable identification of a particular person or organisation.

The *Census and Statistics Act 1905* and the *Privacy Act 1988* require that all information submitted to, or collected by the ABS remain confidential. All ABS staff, including temporary employees, are legally bound never to release personal information to any individual or organisation outside the ABS. In addition, comprehensive security arrangements are implemented in ABS computer systems. These include use of regularly changed passwords, access controls and audit trails.

The authors would like to acknowledge Brendan Kelly, Bradley White, Richard Grant and Paul Campbell for their valuable assistance on this project.

ABBREVIATIONS

ABN	Australian Business Number
ABR	Australian Business Register
ABS	Australian Bureau of Statistics
ACMID	Australian Census and Migrants Integrated Dataset
AMEP	Adult Migration Education Program
ANZSCO	Australian and New Zealand Standard Classification of Occupations
ANZSIC	Australian and New Zealand Standard Industrial Classification
ATO	Australian Taxation Office
DIAC	Department of Immigration and Citizenship
DIBP	Department of Immigration and Border Protection
D-MAC	Deterministic linking Macro
DSS	Department of Social Services
ERP	Estimated Resident Population
ESTFN	Encrypted Scrambled Tax File Number
MUR	Marginal Uniqueness Rate
NMSU	National Migrants Statistics Unit
PAYG	Pay-As-You-Go
PIT	Personal Income Tax
PITMID	Personal Income Tax and Migrants Integrated Dataset
SAS	Statistical Analysis Software
SDB	Settlement Database
TRIPS	Travel and Immigration Processing System

CONTENTS

PERSONAL INCOME TAX AND MIGRANTS INTEGRATED DATASET (PITMID) 2011–12 QUALITY ASSESSMENT

National Migrant Statistics Unit
Australian Bureau of Statistics

ABSTRACT

The Australian Bureau of Statistics (ABS) created the Personal Income Tax and Migrants Integrated Dataset (PITMID) by linking Australian Taxation Office personal income tax records with migrant records from the Australian Government's Settlement Database (SDB). In 2015, PITMID data for 2009–10 and 2010–11 was released in *Personal Income of Migrants, Australia, Experimental* (ABS cat. no. 3418.0). In 2016, this PITMID study was conducted to assess the effects of a change in the linking methodology for the 2011–12 PITMID. The study briefly outlines the original and new linking methodologies and presents the results of the analyses conducted to assess the quality of the 2011–12 PITMID linkage compared with the 2009–10 and 2010–11 PITMID linkages. The new methodology utilising the D-MAC was found to be much quicker to administer and produced high quality results, while enabling comparison between the annual series.

1. INTRODUCTION

The Personal Income Tax and Migrants project is an ongoing project that aims to provide detailed information on the income characteristics of permanent migrants by integrating permanent migrant settlement records from the Australian Government's Settlement Database (SDB) with records from the Australian Taxation Office's (ATO) Personal Income Tax (PIT) data.

In 2013, the ABS created the 2009–10 and 2010–11 Personal Income Tax and Migrants Integrated Datasets (PITMID), by linking permanent and provisional migrant settlement records to personal income taxation records for 2009–10 and 2010–11. The feasibility of linking these datasets and the quality of the resulting datasets were assessed in *Feasibility Study of Linking Migrant Settlement Records to Personal Income Tax Data* (ABS, 2014). Detailed statistics on both datasets were released in 2015 in the publication *Personal Income of Migrants, Australia, Experimental* (ABS, 2015).

This project has three phases: Feasibility, Dissemination and Production. The creation of the 2011–12 PITMID dataset marks the beginning of the Production Phase. For more information on these phases, refer to Appendix A.

When creating the 2011–12 PITMID dataset, it was decided to change the linkage process to utilise the new Deterministic Linking Macro (D-MAC) to improve efficiency.

This paper:

- describes the sources of administrative data for the PITMID (Section 2);
- provides a summary of the linkage process (Section 3);
- analyses the results from the new linkage process and compares to the results obtained when creating the 2009–10 and 2010–11 PITMID datasets (Sections 4 and 5);
- highlights some of the new statistics that are available from the 2011–12 PITMID dataset (Section 6); and
- addresses considerations for future iterations of the project (Section 7).

References to migrants are in the context of those who linked to an available PIT record, indicating that they have submitted a tax return or earned a wage or salary in a job where their employer lodged a Pay-As-You-Go form on their behalf in the financial year. The data has not been calibrated to reflect the total population of migrants who submitted a tax return and thus may not be representative of the whole migrant taxpayer population.

2. THE SOURCES OF ADMINISTRATIVE DATA FOR PITMID

This section provides an overview of the two administrative data sources used to create the 2011–12 PITMID. The first data source is the Australian Government's Settlement Database (SDB) and the second data source is the Australian Taxation Office's (ATO) Personal Income Tax (PIT) unit records.

2.1 Settlement Database (SDB)

The SDB is compiled by the Australian Government from various departmental systems and a number of external sources, including Medicare Australia (DIAC 2013). The Department of Social Services (DSS) has custodianship of the database.

The SDB is a consolidated database of people who have been granted a permanent or a provisional visa (DIAC 2013). The SDB generally excludes temporary visa holders. However, there are some records for people on provisional visas as they transition to permanent residency.

For Settlement visas that were granted onshore (i.e. in Australia), the Arrival Date refers to the latest date of arrival prior to the grant of that visa. For Settlement visas that were granted offshore (i.e. outside of Australia), the Arrival Date refers to the first date of arrival after the visa was granted.

The SDB dataset contained 1,998,473 records of people who were granted a permanent or provisional visa between 1 January 2000 and 6 March 2013. The SDB data was also supplemented by name and address history data from the Department of Immigration and Border Protection (DIBP) Travel and Immigration Processing System (TRIPS).

After analysing the SDB records prior to linking the 2009–10 and 2010–11 files, 225,140 records on the SDB were identified as being from people who were deceased prior to the reference period or aged less than ten years at the start of the reference period. This left 1,773,333 SDB records considered acceptable for linking. Despite assessing all SDB records for the 2011–12 linkage process, analysis variables were only kept for these 1,773,333 records.

For a full list of the linking and analysis variables available on the SDB, see table B.1 in Appendix B.

2.2 Personal Income Tax (PIT)

The PIT unit record data are sourced from ATO taxation returns processed up to 16 months after the end of the financial year (i.e. returns processed up to 31 October 2013 for the financial year ending 30 June 2012). The number of unique PIT records received for the 2011–12 financial year was 12,735,689.

The PIT data is supplied to the Australian Statistician under the *Taxation Administration Act 1953* for the purposes of administering the *Census and Statistics Act 1905*. The data has been collected in compliance with Australian taxation laws. The unit record data was provided to the ABS for a variety of statistical purposes and so was not tailored specifically to this project.

Due to the identifying nature of the data it contains, access to all ATO datasets is strictly regulated by the ATO. Both the ATO and the ABS handle personal information contained in the data in accordance with the Australian Privacy Principles contained in the *Privacy Act 1988*.

According to taxation laws, individuals whose income is below a certain threshold are not required to submit taxation returns. In the 2009–10 and 2010–11 financial years, the tax-free threshold was $6,000 and it remained at this same level for the 2011–12 financial year.

For a full list of the linking and analysis variables available on the PIT, see table B.2 in Appendix B.

2.3 Linking variables

The variables used in the linking process are listed in table 2.1 below.

2.1 Variables used in linking

Variable type	*SDB*	*PIT*
Name information	Given names (Anonymised)	Given name (Anonymised)
	First initial (Anonymised)	First initial (Anonymised)
	Surname (Anonymised)	Surname or family name (Anonymised)
	Alias first name (Anonymised)	Given name (Anonymised)
	Alias surname (Anonymised)	Surname or family name (Anonymised)
Personal characteristics	Date of birth	Date of birth
	Sex	Sex
Address information	Current MB	MB
	Current SA1	SA1
	Current SA2	SA2
	Current SA4	SA4
	Current State	State
	Previous MB	MB
	Previous SA1	SA1
	Previous SA2	SA2
	Previous SA4	SA4
	Previous State	State

An analysis of the linking variables indicated that there was a very low rate of missing data.

3. THE LINKING PROCESS

Before the commencement of the 2011–12 PITMID linkage process, records that had been linked in 2009–10 and 2010–11 (951,234) were removed along with 388 records for deceased migrants and records for those people who were less than 10 years of age at the beginning of the reference period, 1 July 2011 (174,189 SDB records and 14,026 PIT records). This resulted in a final file of 872,662 SDB migrant records for linking to the 11,880,880 PIT records in 2011–12.

The strategy used for 2011–12 PITMID involved a number of linking runs. Initial linking runs required exact matches on Name and Address information at the Meshblock and SA1 level. Later runs allowed for more variation in the linking variables, such as Address at SA2 level and Jaro-Winkler scores[2] of at least 0.91 for Name.

See Appendix C for details of the linkage runs.

3.1 Changes in methodology introduced for the 2011–12 PITMID

The linkage strategy for 2011–12 PITMID differed slightly from the strategy used for the 2009–10 and 2010–11 linkage. Most significant was the implementation of the linking strategy using the new Deterministic linking Macro (D-MAC).

The 2009–10 and 2010–11 PITMID were created using a combination of a Deterministic[3] and Probabilistic[4] linking approaches to link the SDB and PIT datasets.[5]

For 2011–12 PITMID, the D-MAC provided a purely deterministic approach. The following points were considered when making this change:

- The limited number of linking variables;
- The high quality of the available linking variables;
- The unknown overlap between the populations on the two datasets; and
- The D-MAC process would be quicker and more efficient.

2 The Jaro-Winkler distance is a measure of similarity between two strings. The higher the Jaro-Winkler distance for two strings is, the more similar the strings are. The score is normalised such that 0 equates to no similarity and 1 is an exact match.

3 Deterministic linking links two records together if all linking fields are identical and only one such record meets this criterion.

4 Probabilistic linking links two variables together using link weights to calculate the probability that two given records refer to the same entity.

5 For more details see *Feasibility Study of Linking Migrant Settlement Records to Personal Income Tax Data* (ABS, 2014).

3.1.1 Deterministic linking macro (D-MAC)

Deterministic linking compares two records on a set of variables. If all variables agree, they are considered a link. Multiple iterations, or passes, can be undertaken using different sets of variables on which to link. Typically, deterministic linkage begins by using very stringent matching rules (where the records pairs need to agree exactly on as many linking fields as possible). As some matches will not agree on all linking variables, subsequent deterministic passes typically relax linking conditions by removing a variable from the set of comparison variables. However, in order to establish a unique agreement, all sets of variables used in deterministic passes must be strongly identifying. The strength of deterministic linking is that it can quickly locate the high quality matches in a dataset.

The D-MAC is a SAS macro which was developed by the ABS for deterministically linking administrative datasets. D-MAC identifies unique links between two datasets on a given set of conditions and uses two measures, the Marginal Uniqueness Rate (MUR) and the Duplicate Rate as a measure of the quality and accuracy of linked pairs.

The Duplicate rate is the number of records on Dataset 1 (i.e. the SDB) that agree with at least 2 records on Dataset 2 (i.e. the PIT), divided by the number of records on Dataset 1 that agree with at least 1 record on Dataset 2. A low Duplicate rate means a high proportion of records uniquely agree (i.e. match only one record on Dataset 2).

$$\text{Duplicate rate} = \frac{\#\text{ records on Dataset 1 that agree with at least two records on Dataset 2}}{\#\text{ records on Dataset 1 that agree with at least 1 record on Dataset 2}}.$$

The MUR is a similar concept, but is a pass-dependent measure. It is best defined as the number of best links[6] in a pass that are unique, divided by the number of best links in a pass. If a record links uniquely in a pass, but is linked more convincingly in another pass (i.e. in a pass where all records are uniquely linked) it will not contribute to the MUR of the pass.

$$\text{MUR}_{\text{Pass N}} = \frac{\#\text{ Records on Dataset 1 that link best in Pass N that are unique}}{\#\text{ Records on Dataset 1 that link best in Pass N}}.$$

For the 2011–12 PITMID, cut-off values for MUR were identified to collate quality linked pairs. For further information, see Section 4.

6 The best link is determined after consideration of all of the alternative links generated. The best link usually depends upon which quality measure is used to rank the generated links. In this instance, the quality measure utilised is the Duplicate Rate. Therefore, the best link is the link with the lowest Duplicate Rate in a given pass.

The D-MAC was also applied to the previous 2009–10 and 2010–11 PITMID datasets to enable a comparison of the new linkage results with the previous linkage results to assess the effectiveness of the new method.

During the linking it was discovered that records with very common first names can inflate the Duplicate Rate in D-MAC and decrease the MUR. This can result in some very high quality links with less common first names being assigned a low MUR, if they are assigned at all, and hence not be accepted as valid links. This issue was addressed by recoding First Name into three mutually exclusive linking variables, i.e. Common First Name, Uncommon First Name and Rare First Name. Only one of these variables is permitted in each pass and each record can have a non-missing value for only one of the three linking variables. Grouping the names in this way allows for a greater understanding of the quality and relative uniqueness of the links generated. This is reflected in the Duplicate Rate and MUR and provides a better overall quality estimate.

3.1.2 Name anonymisation

Given name and surname information was encrypted for the 2011–12 PITMID linking using an anonymisation method that was developed by the ABS. The name information on both the SDB and the PIT datasets is of very high quality. If name information is of high quality, it is generally accepted that there will be no loss in quality when using anonymised name rather than the actual name as a linking variable.

3.1.3 Spouse linking

Spouse information was not available on 2011–12 PIT file. Therefore passes that used this information could not be replicated.

3.1.4 Forwarding address

A forwarding address on the SDB could not be ascertained due to the PIT data cut-off date occurring after the SDB extraction date. For this reason, only current address and previous address on the SDB address history file were used.

4. LINKAGE RESULTS

After the linkage process was completed, an appropriate cut-off value for Marginal Uniqueness Rate (MUR) was determined and the following results were obtained.

4.1 Linkage results

Run	*Links generated*	*MUR cut-off*	*Links with MUR ≥ cut-off (retained)*	*Links retained*
	No.		No.	%
1	49,806	0.99	49,689	99.77
2	27,302	0.99	27,206	99.65
3	7,455	0.98	7,152	95.94
4	110,837	0.97	52,228	47.12
5	1	0.97	1	100.00
6	17,246	0.97 with conditions	722	4.19
Total	**212,647**		**136,998**	**64.48**

Run 6 utilised a 'Drop one' strategy (i.e. one variable was dropped from the linkage pass to allow for missing responses). Only links meeting certain conditions were accepted. The links had to have an MUR of at least 0.97 and the following conditions were applied to each dropped variable.

- If Given name had been dropped, accept if the records matched at MB or SA1 level;
- If Sex had been dropped, accept if the records matched at MB or SA1 level; and
- If Date of birth had been dropped, accept if the records matched at MB or SA1 level.

No links were retained if Surname had been dropped.

4.2 Number of linked and unlinked SDB records by financial year

	No.	%
Linked in 2009–10 or 2010–11	951,234	53.64
Linked in 2011–12	136,998	7.73
Total linked	1,088,232	61.37
Unlinked	685,101	38.63
Total	**1,773,333**	**100.00**

A total of 212,647 SDB records linked to a PIT record, with 136,998 of these links meeting the MUR cut-off value to be considered true links in 2011–12. This raises the total number of SDB records that have linked to a PIT record (2009–10, 2010–11 and 2011–12) to 1,088,349. Most of these links were established in 2009–10 and 2010–11 (87%) while the remaining 13% were linked during the 2011–12 PITMID linkage.

Adding these new links to the links that had already been generated in the previous linking runs, the results reported in table 4.2 were obtained and the linkage rate increased from 54% to 61%.

4.1 Linkage rates by year of arrival

Table 4.3 provides the linkage rates for each PITMID created so far by the migrant taxpayers' year of arrival.

Year of arrival is calculated differently depending on whether the visa application was lodged onshore or offshore. If a visa application was lodged offshore, year of arrival refers to the year that the migrant first arrived after the grant of that visa. If a visa application was lodged onshore, year of arrival is the most recent year the migrant arrived in Australia prior to the grant of that visa.

Persons who have a year of arrival after the reference period may have earned an Australian taxable income from overseas. Alternatively the migrant may have:

- Initially arrived in Australia on a temporary visa and submitted a tax return for 2011–12;
- Left Australia;
- Applied for a permanent visa;
- Arrived back in Australia after the reference period.

4.3 Number of linked and unlinked SDB records by Year of arrival

	Linked records				
Year of arrival	*2009–10 or 2010–11*	*2011–12*	*Total*	*Unlinked records*	*Total*
	No.	No.	No.	No.	No.
2000	40,749	5,025	45,774	37,833	83,607
2001	46,297	5,210	51,507	38,456	89,963
2002	52,223	5,812	58,035	40,078	98,113
2003	67,008	7,074	74,082	45,886	119,968
2004	76,008	7,285	83,293	48,475	131,768
2005	87,921	7,091	95,012	48,502	143,514
2006	99,481	7,758	107,239	48,243	155,482
2007	107,933	8,710	116,643	48,279	164,922
2008	119,534	10,792	130,326	55,229	185,555
2009	111,358	13,602	124,960	57,672	182,632
2010	79,980	16,059	96,039	53,531	149,570
2011	41,626	30,729	72,355	65,922	138,277
2012	19,776	11,535	31,311	84,878	116,189
2013	1,340	316	1,656	12,117	13,773
Total	**951,234**	**136,998**	**1,088,232**	**685,101**	**1,773,333**

Those migrant taxpayers who linked for the first time in 2011–12 but arrived in Australia prior to the 2009–10 or 2010–11 financial years, may not have linked previously for a number of reasons including;

- Their income in 2009–10 and 2010–11 fell below the tax free threshold of $6,000 and they were not obliged to submit a taxation return for those financial years;
- They have not yet lodged their taxation return for 2009–10 or 2010–11; or
- They did submit a taxation return in 2009–10 or 2010–11 but their name and address information was not of sufficient quality to link with a settlement record.

While the majority of SDB records were linked to the 2009–10 and 2010–11 PIT records (the first iteration), linking the remaining records to the 2011–12 PIT records (the second iteration) had a significant impact when it came to linking records that had a more recent year of arrival. For example, 40% of records for people who arrived from 2011 to 2013 linked in the second iteration.

4.2 Match rate analysis

In order to assess the success of the PITMID linkage compared with other sources, a match rate analysis was conducted. Given the linkage rate for this project was not expected to reach 100%, match rate can be used to give an indication of the quality of the linkage results.

The match rate (also referred to as Recall, or Recall rate) is defined as the number of true links over the expected total number of records that appear on both datasets. Since this number is unknown, it has to be estimated.

As was the case with the 2009–10 and 2010–11 results, the match rate estimate is calculated as:

$$\text{Expected match rate} = \frac{\text{Number of SDB records correctly linked}}{\text{Expected number of SDB matches}}$$

In order to identify the population most likely to have submitted a tax return, the SDB population was limited to those individuals who:

- were aged 15 to 64 at the beginning of the reference period (01/07/2011);
- had a permanent (Skill, Family, Humanitarian or Other permanent) or provisional visa;
- had a visa grant date prior to the beginning of the reference period (01/07/2011); and
- had not departed Australia prior to the reference period according to their TRIPS data.

Individuals who met all of these conditions were considered to be 'eligible taxpayers' for the following analysis.

Table 4.4 shows the number of migrant taxpayers in 2011–12 who were linked by their eligibility status.

4.4 Number of linked and unlinked 2011–12 records by eligibility status

	Linked records	*Unlinked records*	*Total*
	No.	No.	No.
Eligible taxpayers	798,921 c	249,862	1,048,783 a
Not eligible taxpayers	289,311	435,239	724,550
Total	**1,088,232**	**685,101**	**1,773,333**

a = Number of eligible taxpayers. This figure is used in match rate analysis in table 4.6

c = Number of actual links. This figure is used in match rate analysis in table 4.6

Table 4.5 shows population figures for the Australian Population aged 15–64 years as at 30 June 2012, according to the Australian Demographic Statistics publication (ABS, 2016a), as well as the number of Australian taxpayers for the 2011–12 financial year according to the ATO's 2011–12 Taxation Statistics publication (ATO, 2014). These figures can be used to estimate the actual number of true links for the 2011–12 match rate analysis.

4.5 Proportion of persons aged 15–64 submitting a tax return for the 2011–12 financial year

	Number of records
Australian Demographic Statistics – 2011–12 – Persons aged 15–64 years	15,209,716
ATO's Taxation Statistics – 2011–12 – Persons aged 0–64 years	11,530,605
Proportion of persons submitting a tax return in 2011–12	75.81 b

b = Proportion of the Australian population submitting a tax return. This figure is used in match rate analysis in table 4.6

The 2011–12 ATO Taxation Statistics data relates to persons aged "0–64 years". Those aged "15–17 years" could not be identified in the 2011–12 ATO Taxation Statistics data due to the younger age group being defined as persons aged "Less than 18 years".

Table 4.5 includes ATO data for those aged "0–64 years" with 172,000 of these taxpayers aged "Less than 18 years "of age. For the purposes of this analysis and to enable comparison, it is assumed that the likelihood of being a taxpayer increases with age and that a significant number of these taxpayers would be aged "15–17 years". However, using data for taxpayers aged "0–64 years" instead of taxpayers aged "15–64 years" leads to a higher proportion of taxpayers, resulting in a lower match rate, giving a conservative result.

The Match rate is premised on two assumptions: first, that all links are true links and second, that recent migrants are just as likely to submit a tax return as the rest of the population. Given linking was conducted using name and address and only high quality pairs were linked, we expect the first assumption to hold. The second assumption appears not to hold, however, as a match rate greater than 100% is not possible in theory. The match rate of 100.5% indicates that migrants aged 15–64 years are more likely than the Australian population (of the same age range) to submit a tax return. Nonetheless, the high match rate here suggests a high quality linked file with few missed matches.

Table 4.6 below shows the calculation of the match rate.

4.6 Match rate analysis for the 2011–12 financial year

	Formula	*Calculation*
Number of Eligible taxpayers	a	1,048,783
Proportion of persons submitting a tax return	b	75.81%
Expected number of matches	a×b	1,048,783 × 75.81% = 795,082.4
Number of actual links	c	798,921
Match rate	c/(a×b)	798,921 / 795,082.4 = 100.5%

4.3 The comparison of the linkage processes

To assess the accuracy of the D-MAC linking methodology, it was decided to recreate the 2009–10 and 2010–11 PITMID linkage using D-MAC and compare the linkage results with those obtained during the first iteration. The SDB and 2009–10 and 2010–11 PIT datasets were prepared in the same way as the 2011–12 files, including encrypting name data and considering "current address" as the most recent address recorded on the address history file.

Table 4.7 presents the D-MAC Linking results for the 2009–10 and 2010–11 financial year data.

4.7 D-MAC Linking results for 2009–10 and 2010–11 SDB to PIT linkage

Year of arrival	*Passes*	*Links*	*MUR cut-off*	*Links retained*	*Links retained*
	No.	No.		No.	%
1	34	571,714	0.99	571,714	100.00
2	78	47,222	0.99	42,683	90.39
3	214	36,743	0.98	36,743	100.00
4	420	338,623	0.97	271,378	80.14
5	6	2	0.97	0	0
6	166	44,659	0.97 with conditions	14,890	33.34
Total	**918**	**1,038,963**		**937,408**	**90.23**

Table 4.8 shows the number of 2009–10 and 2010–11 records linked in the first iteration linkage that were also linked with D-MAC. The results are also shown for those records that were linked with D-MAC, but were not linked previously.

4.8 D-MAC Linking results for 2009–10 and 2010–11 SDB to PIT linkage compared with original linkage results

	Linked without using spouse information		*Linked using spouse information*		*Total*	
Result	*No.*	*%*	*No.*	*%*	*No.*	*%*
Records linked in the previous 2009–10 and 2010–11 linkage process						
Linked with DMAC						
To the same PIT record	862,867	48.66	30,825	1.74	893,692	50.39
To a different PIT record	649	0.04	4	0.00	653	0.04
Total	863,516	48.69	30,829	1.74	894,345	50.43
Not linked with DMAC	56,378	3.18	511	0.03	56,889	3.21
Total	919,894	51.87	31,340	1.77	951,234	53.64
Records not linked in the previous 2009–10 and 2010–11 linkage process						
Linked with DMAC	43,063	2.43	0	0.00	43,063	2.43
Not linked with DMAC	779,036	43.93	0	0.00	779,036	43.93
Total	822,099	46.36	0	0.00	822,099	46.36
Total	**1,742,072**	**98.23**	**31,340**	**1.77**	**1,773,333**	**100.00**

Almost 95% of all SDB records achieved the same results irrespective of linkage method i.e. either linked to the same PIT record as in the previous 2009–10 and 2010–11 linkage process, or did not link.

The linking with D-MAC without spouse information resulted in 98% of the records linked in the first iteration linkage (with spouse information) being linked to the same PIT record with D-MAC. This indicated that spouse information had a negligible impact upon the linking outcome.

A sizable number (14,200) of the records that linked in the previous iteration (but did not link with D-MAC) came from a single pass. This pass was limited to females only and did not use surname as a linking variable. The main purpose of this pass was to account for changes in surname for females (e.g. due to marriage or divorce). By contrast, the D-MAC linkage strategy did not retain any links generated in passes that did not link on surname. Links generated in Run 6 that dropped surname were not retained.

Similar results are obtained with D-MAC compared with the previous linkage methodology. The first iteration linkage process resulted in a linkage rate of 53.6%, whilst the new methodology with D-MAC resulted in a linkage rate of 52.9%.

The analysis demonstrates that the new linking method with D-MAC produces a dataset comparable with the one produced during the first iteration. Given the dataset produced during the first iteration was assessed to be of high quality, it is safe to assume that the dataset produced using D-MAC is also of high quality.

5. ANALYSIS DATASETS

5.1 PIT data record counts

After the linking process was completed, Client and PAYG dataset information was provided only for those records that linked to an SDB record for each reference period.

The PAYG dataset contained duplicate records where an individual worked in more than one job in the financial year. The number of PAYG records for each individual on the dataset ranged from 1 to greater than 20.

Table 5.1 shows the PAYG and Client data record counts for 2009–10, 2010–11 and 2011–12.

5.1 Client dataset and PAYG dataset record counts for linked records Client dataset

	2009–10	*2010–11*	*2011–12*	*% Increase from 2009–10 to 2010–11*	*% Increase from 2010–11 to 2011–12*
ATO Client data records	812,482	888,542	977,777	9.36	10.04
ATO PAYG records	1,032,913	1,158,611	1,494,637	12.17	29.00

The increase in the number of Client data records from 2010–11 to 2011–12 is similar to the increase from 2009–10 to 2010–11.

The increase in the number of PAYG records from 2010–11 to 2011–12 is due to the fact that for 2009–10 and 2010–11, PAYG records were requested only for linked records on the Client dataset. For 2011–12, PAYG records were requested for all records (including those not on the Client dataset) that linked to PAYG records. This change was implemented to determine if records that appeared on the PAYG dataset and not on the Client dataset were mainly individuals with taxable incomes below the tax-free threshold.

After removal of records on the PAYG dataset with no income information, there were almost 1.25 million records on the PAYG dataset relating to 841,233 unique persons. In total, there were 1,088,227 unique persons on at least one of the datasets.

Table 5.2 shows the records present on the 2011–12 Client dataset and the Individual PAYG dataset.

5.2 Record counts for persons on Client dataset and PAYG dataset, 2011–12

	On PAYG dataset	*Not on PAYG dataset*	*Total*
On Client dataset	810,234	167,543	977,777
Not on Client dataset	30,999	79,451	110,450
Total	**841,233**	**246,994**	**1,088,227**

Of the 30,999 individuals on the PAYG dataset not present on the Client dataset, almost 35% had a total for "Gross payments" less than $6,000 on the PAYG dataset. As such, these individuals would not have been obliged to lodge a taxation return in 2011–12.

Around two-thirds of individuals who are not present on the Client file have incomes above the tax-free threshold. These individuals may not be present on the Client dataset due to:

- Not having submitted a tax return in the financial year;
- Not having had their tax return processed by the ATO by the cut-off date (31 October 2013).

5.2 Comparison to 2009–10 and 2010–11 PITMID files

Records on the PITMID were limited to those who had reported either a value for at least one income source or a value for "Total deductions". The total number of records on the 2011–12 PITMID file is 1,002,179.

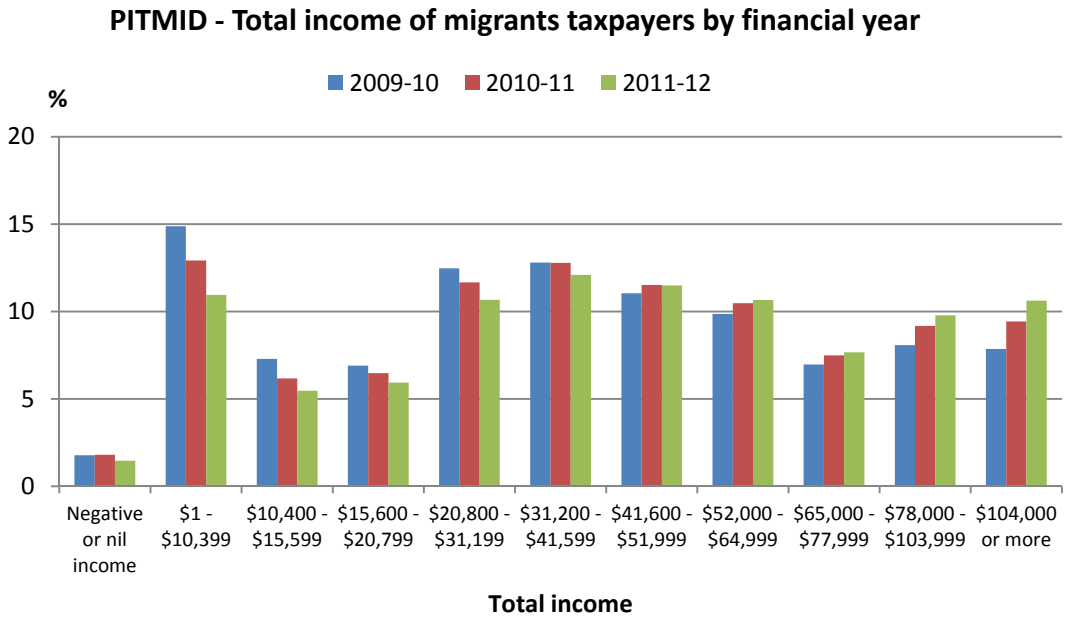
With the creation of the 2011–12 PITMID, there are now potentially three data points for records that have been present in all three PITMID files. As such, there are over 700,000 records that are present in all three datasets, representing those migrants who reported income and/or deductions in all three financial years.

5.3 Number of records on 2009–10, 2010–11 and 2011–12 PITMID files

	Number of records
Records in all three financial years (2009–10, 2010–11 and 2011–12 PITMID)	708,863
Records in both 2009–10 and 2010–11 PITMID	27,858
Records in both 2009–10 and 2011–12 PITMID	28,681
Records in both 2010–11 and 2011–12 PITMID	122,630
Records only in 2009–10 PITMID	30,333
Records only in 2010–11 PITMID	11,706
Records only in 2011–12 PITMID	142,005
Total 2009–10 PITMID records	795,732
Total 2010–11 PITMID records	871,054
Total 2011–12 PITMID records	1,002,179

Figure 5.4 shows the proportion of migrant taxpayers by total income ranges from PITMID for the three financial years. Overall there is a similar pattern of income. For 2011–12 the proportions are slightly smaller for the lower income earners (who reported up to $41,599), which is to be expected as income levels tend to increase over time. There is also a slight incremental increase in the proportion of migrants in the higher income ranges ($52,000 or more) for each financial year.

5.4 Total income, 2009–10, 2010–11 and 2011–12 PITMID files



6. NEW DATA FOR 2011–12

6.1 PAYG – Concurrent and consecutive job holders

The inclusion of data items recording the start and end date for the payment period on the PAYG dataset helps determine which jobs held by a taxpayer (with multiple records on the PAYG dataset) are held concurrently or consecutively. Holding concurrent jobs enables us to identify those migrant taxpayers who are 'multiple jobholders'.

Over two-thirds of migrants with a record on the PAYG dataset held only one job in the financial year. So, for the purposes of this analysis on those who held more than one job in 2011–12, they have been excluded. For 381 records, it was not possible to identify the start and end dates accurately due to inconsistencies in the information provided. These records were coded to "Inadequately described".

Table 6.1 shows that of the migrant taxpayers who held multiple jobs, just over half held more than one job concurrently in the 2011–12 financial year. Of those migrants who held two jobs, about two-thirds held them concurrently and a third consecutively. The situation is reversed for those migrant taxpayers who held three or four jobs. Almost two thirds held their jobs consecutively while just over a third held them concurrently in 2011–12.

6.1 PAYG Data, Proportion of migrant taxpayers with more than one job in 2011–12 by whether jobs held concurrently or consecutively

	Migrants who held concurrent jobs	*Migrants who only held consecutive jobs*	*Inadequately described*	*Total*
	%	%	%	%
Total jobs on PAYG file				
2 jobs	65.18	34.62	0.20	100.00
3–4 jobs	35.38	64.60	0.02	100.00
5 or more jobs	14.19	85.80	0.01	100.00
Total	**54.75**	**45.11**	**0.14**	**100.00**

6.2 Industry of own unincorporated business

When investigating the 2009–10 and 2010–11 PITMID data, it was discovered that a higher proportion of migrants with a Humanitarian visa reported income from Own unincorporated businesses. This led to increased interest in the industries pertaining to this migrant business income. Table 6.2 shows migrant taxpayers with business income by the industry division of the unincorporated business by visa stream for 2011–12 from the Business Income Tax (BIT) dataset.

6.2 Migrants who reported business income, By the industry division of the unincorporated business by visa stream

Industry of own unincorporated business	*Skill*	*Family*	*Humanitarian*	*Provisional*	*Total (a)*
	No.	No.	No.	No.	No.
Agriculture, Forestry & Fishing	574	345	69	15	1,004
Mining	26	12	..	0	39
Manufacturing	1,029	835	105	75	2,046
Electricity, Gas, Water & Waste Services	71	30	..	15	123
Construction	6,573	5,840	2,467	231	15,131
Wholesale Trade	724	367	43	41	1,177
Retail Trade	2,747	1,609	213	193	4,766
Accommodation & Food Services	1,196	926	117	63	2,307
Transport, Postal & Warehousing	7,448	2,228	1,088	2,252	13,024
Information Media & Telecommunications	635	405	16	45	1,102
Financial & Insurance Services	1,589	474	24	16	2,103
Rental, Hiring & Real Estate Services	696	312	22	33	1,064
Professional, Scientific & Technical Services	8,638	3,684	260	366	12,962
Administrative & Support Services	4,463	2,957	539	940	8,907
Public Administration & Safety	368	127	57	111	664
Education & Training	1,620	910	71	67	2,672
Health Care & Social Assistance	6,091	2,648	1,042	110	9,898
Arts & Recreation Services	978	1,090	22	44	2,137
Other Services	3,215	2,487	403	202	6,309
Total (b)	72,557	41,099	8,552	6,565	128,904
No business income	538,619	247,871	40,843	45,131	873,275
Total	611,176	288,970	49,395	51,696	1,002,179

(a) Includes Visa "Other permanent" and "Unknown".

(b) Includes Industry of business "Inadequately described".

.. Not available for publication.

Source: Business Income Tax Dataset, 2011–12

More than one-third of migrants with a provisional visa who report business income in 2011–12 own a business in the Transport, Postal and Warehousing industry. Humanitarian and Family migrants were most likely to own a business in the Construction industry, at 29% and 14% respectively. Most migrants in the Skill stream were in the "Professional, Scientific and Technical Services" at 12%.

6.3 Citizenship status and last visa held

The response for "Visa subclass" is the last visa the migrant was granted. This may not be the same as the visa they currently hold as they may have been granted Australian citizenship.

The date Australian Citizenship was conferred from the SDB can tell us whether or not a migrant had obtained citizenship before the end of the 2011–12 reference period.

With the exception of Other permanent migrants, migrant taxpayers who became Australian citizens before the end of the 2011–12 reference period had higher median total income than those who were not Australian citizens, irrespective of their visa stream.

6.3 Median total income, By Citizenship status and Visa stream

Visa stream	*Australian citizen prior to end of 2011–12 financial year*	*Not an Australian citizen prior to end of 2011–12 financial year*	*Total*
	$	$	$
Skill	56,245	48,142	51,205
Family	39,017	33,675	35,352
Humanitarian	31,291	23,980	28,432
Other permanent	43,009	43,421	43,136
Provisional	..	31,972	31,972
Total	**48,677**	**40,389**	**43,142**

7. CONSIDERATIONS FOR FUTURE ITERATIONS OF PITMID

7.1 Increase of the tax free threshold in 2012–13

For the 2000–01 financial year, the tax free threshold increased from $5,400 to $6,000 and remained at this level until the 2011–12 financial year. In 2012–13, the tax free threshold increased to $18,200. This large increase may have a significant effect on the number of people submitting tax returns in subsequent years because those persons with a taxable income between $6,001 and $18,200 will no longer be required to submit a taxation return. The potential drop in the number of taxpayers and change to median income will mean that any comparisons between the 2012–13 PITMID and previous years will need to account for this change.

7.2 Forwarding address

The next extract of the SDB for the 2012–13 PITMID will have a cut-off date of 1 July 2016. The extract will provide "current address" and "previous address" as well as enable the calculation of a "forwarding address". This is consistent with the 2009–10 and 2010–11 PITMID linking (14,568 records were linked using "forwarding address").

The D-MAC also calculated "current", "previous" and "forwarding" address for the 2009–10, 2010–11 and 2011–12 PIT files. However, adding forwarding address made very little difference to the linking results, with only 391 additional records linked. While forwarding address will be available for the 2012–13 and 2013–14 PIT file, it will not be available for the last two years (2014–15 and 2015–16). Given the limited gains achieved with forwarding address, it is not envisaged that the absence of this data item will have any significant impact on the linkage.

8. CONCLUSION

Overall, linking the 2011–12 PIT data to the SDB records using D-MAC was successful. An extra 137,000 linked record pairs were identified and were added to the 951,000 links generated when linking the 2009–10 and 2010–11 datasets. The utilisation of D-MAC meant that the data was linked efficiently and only unique record pairs were identified, eliminating the need for extensive clerical review and greatly decreasing the amount of time needed to link the datasets.

The new methodology utilising the D-MAC was found to be much quicker to administer and produced results similar to the previous linking method. Given that the previous linking method was assessed to have produced high quality results, it is reasonable to infer that the D-MAC methodology also produces high quality results. The similarity of the results enables comparison between the annual series (i.e. 2009–10, 2010–11 and 2011–12). When using D-MAC to link the SDB to the 2009–10 and 2010–11 PIT datasets, over 95% of the SDB records received exactly the same link result as they did when linked during the original linking process. For this reason, the links generated for 2009–10 and 2010–11 were retained for the 2011–12 linkage process.

Due to the similarity of the results obtained by both processes, as well as the expedited nature of linking using D-MAC, the PITMID Project will continue to use the D-MAC for linking in future. However, some minor changes to the linking process for 2012–13 may be considered. These include:

- Remove passes using "First initial" in Run 4. When looking at such broad geography as SA4 and State, not enough information can be retained using just first initial.
- Remove Run 5 as it did not generate many links.
- Limit passes in Run 6 to just Meshblock and SA1 level geography as SA2 is considered too broad when dropping one of the linking variables.
- Consider retaining links generated when Surname is dropped, but limit to females, in order to identify women who have changed their surname due to marriage or divorce.

Future changes in taxation laws may occur and would result in changes to the questions asked of taxpayers on the taxation return. This may lead to changes in the PIT data available for PITMID. However, any new PIT variables could also result in new income statistics for permanent migrants and those with a provisional visa.

The D-MAC is becoming the preferred linking method for other important ABS data integration projects. Given the unique nature of the PIT data items and the relatively few demographic data items available on the dataset for linking, any dataset being integrated with the PIT data will most likely need to contain detailed personal information such as name, date of birth and address. These demographic data items were essential for successful linkage enabling high quality links to be established for PITMID.

REFERENCES

Australian Bureau of Statistics (2014) "Feasibility Study of Linking Migrant Settlement Records to Personal Income Tax Data", *Methodology Research Papers*, cat. no. 1351.0.55.051, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.051 >

—— (2015) *Personal Income of Migrants, Australia, Experimental, 2010–11*, cat. no. 3418.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/3418.0 >

—— (2016a) *Australian Demographic Statistics, Sep 2015*, cat. no. 3101.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0 >

—— (2016b) *Information Paper: Transforming Statistics for the Future*, cat. no. 1015.0, ABS, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1015.0 >

Australian Taxation Office (2014) *Taxation Statistics, 2011–12*, ATO, Canberra.
<https://www.ato.gov.au/About-ATO/Research-and-statistics/In-detail/Taxation-statistics/Taxation-statistics-2011-12/>

Department of Immigration and Citizenship (2013) *Settlement Reporting Facility SRF Data Dictionary (External), Version 2*, DIAC, Canberra.
<http://www.immi.gov.au/living-in-australia/delivering-assistance/settlement-reporting-facility/pdf/ext-data-dictionary.pdf>

Richter, K.; Saher, G. and Campbell, P. (2013) "Assessing the Quality of Linking Migrants Settlement Records to 2011 Census Data", *Methodology Research Papers*, cat. no. 1351.0.55.043, Australian Bureau of Statistics, Canberra.
< http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.043 >

All URLs last viewed on Monday 19 September 2016.

APPENDIXES

A. PERSONAL INCOME TAX AND MIGRANTS PROJECT: PHASES

Phase 1: Feasibility

The Migrant PIT Linkage Feasibility Study aimed to assess the quality of linking PIT records for the financial years 2009–10 and 2010–11 to the individual records of permanent migrants (who arrived on or after 1 January 2000 on the SDB) without the use of a unique record identifier. This linkage was deemed feasible and the resultant linked PITMID experimental datasets for 2009–10 and 2010–11 were created. For more detail see *Feasibility Study of Linking Migrant Settlement Records to Personal income Tax Data* (ABS, 2014). This phase is completed.

Phase 2: Dissemination

The experimental linked dataset created during Phase 1 was considered to be of sufficient quality and usefulness for dissemination. Aggregate data from the experimental PITMID were disseminated in the 2009–10 and 2010–11 statistical releases of *Personal Income of Migrants, Australia, Experimental* (ABS, 2015) in September and December 2015 with finer level data available via customised consultancies on request. This phase is completed.

Phase 3: Production

The current phase. As the custodians, stakeholders and clients were satisfied with the quality, usefulness and protections provided to the PITMID data released during Phase 2, a process is being established to produce PITMID and disseminate data on an annual basis. The 2011–12 PITMID statistical release is expected in October 2016.

B. DATASET STRUCTURE AND AVAILABLE DATA ITEMS

B.1 Settlement Database

The Settlement Database data was sourced from a variety of different databases contained within the Department of Immigration and Border Protection, in particular the Settlement Database (SDB) and the Travel and Immigration Processing System (TRIPS). The data was provided to the ABS in several datasets that were merged together using a person level identifier common to all datasets.

Dataset 1: SDB Client information

The SDB Client file contains the records of 1,998,473 persons who were granted permanent or provisional visas to stay in Australia between 1 January 2000 and 6 March 2013. The records contain demographic information including name, date of birth, sex, country of birth, Australian citizenship, foreign citizenship (country) and information pertaining to the migration event including visa subclass, applicant status (primary/secondary) and location of application (onshore/offshore).

Dataset 2: TRIPS Name history information

A TRIPS Name history dataset was used to evaluate and repair name fields on the SDB Client file. Names on the dataset are correct to the date of extraction and names are updated from a variety of sources including Medicare, the Adult Migration Education Program (AMEP) and manual updates. Individuals who change their name, e.g. women who change their family name when marrying or divorcing, present an issue for linking. The TRIPS name history information helps to address this issue. This dataset contained 407,666 records relating to the first names and surnames of 347,086 persons. The number of name records per person ranged from 1 to more than 10.

Dataset 3: TRIPS Address history information

This dataset provides a complete address history for all records on the SDB file. The address information on the SDB is updated monthly via an administrative process run from Medicare Australia. Therefore, if an individual does not notify Medicare of a change of address, then the SDB record is not updated unless notification is received from the client directly or an update is received from the AMEP.

The address history dataset contained 4,168,920 address records corresponding to the 1,998,473 persons on the SDB Client file (Dataset 1). The records covered the period from 1 January 2000 to 6 March 2013. The number of addresses per person ranged from 1 to more than 20.

Dataset 4: TRIPS arrival and departure information

TRIPS data also recorded the last recorded movement of a migrant, either an arrival to or departure from Australia, and the date on which this movement occurred. This information was used in the Match Rate Analysis in Section 4.2 to identify migrants who had departed Australia prior to the beginning of the reference period and had not returned, and was therefore not retained on the analysis dataset.

The dataset used was originally provided for use in the creation of the *Australian Census and Migrants Integrated Dataset (ACMID), 2011*. The dataset contained a person's last movement direction (arrival or departure) and last movement date information as at Census night (9 August 2011). Permission was granted from DIBP for this data to be utilised for the PITMID feasibility study. It is important to note that the dataset only contained records up to Census night (9 August 2011) rather than up to the date of the SDB extract for this project (6 March 2013).

B.2 Personal Income Tax, 2011–12

The Personal Income Tax data is provided to the ABS by the Australian Taxation Office. In order to maintain the separation principle, identifying information such as name and address are stored on a separate dataset from the analysis variables.

The Personal Income Tax data is in three datasets, the first containing the linking variables, the second analysis variables at the person level and the third analysis variables at the job level.

Dataset 1: Name and address register

The ATO Name and Address register file contained demographic information of persons who have submitted a tax return. Information in the Name and Address register was used for linking and also included sex and date of birth. The number of records received for the 2011–12 financial year was 15,897,252. There were some duplicated instances of tax file number. Duplicates were retained on the file as other information, such as name, were known to vary between the records. The number of unique person records received for linking the 2011–12 financial year was 12,735,689.

Dataset 2: Client dataset

The ATO Client dataset contained a range of variables. These included:

- Wage and salary income;
- Own unincorporated business income (both primary and non-primary production);
- Investment income (including rental income);
- Superannuation and annuity income;
- Government pensions and allowances;

- Taxable income;
- Foreign sources of income; and
- Occupation in main job.

This information was only suppled after linking was completed and only for linked records.

A typical person does not need to fill out every data item in a tax return. As a result, most records have most fields blank in this dataset.

Dataset 3: Individual Pay As You Go (PAYG) dataset

The ATO PAYG dataset contained employer submitted wage and salary information for individuals. Analysis variables available on the PAYG dataset include:

- Gross payments amount;
- Tax withheld;
- Days in job; and
- Average pay per day.

The PAYG file contains a record for every job held by a person throughout the financial year. For 2009–10 and 2010–11 PITMID a count of the number of jobs a person held during the year was obtained from the PAYG file, but it could not be determined whether the jobs were held consecutively or concurrently. For the 2011–12 PITMID, job start and end dates are available from the Individual PAYG dataset (enabling the 'Days in job' variable to be derived). This will enable the identification of concurrent and/or consecutive jobs held by individuals.

Dataset 4: Business Income Tax dataset

The majority of the information on the Client dataset is obtained from the tax return submitted by the individual. However, Industry of own unincorporated business was requested and was merged onto the Client dataset from the Business Income Tax dataset.

Dataset 5: Australian Business Register

The Australian Business Register (ABR) stores details about businesses by ABN. This information is used to identify:

- Industry of employer; and
- Economic sector of employer.

B.1 Data items available on the SDB linking and analysis datasets

Variable name	*Source*
Linking dataset	
Given name	SDB Client information
Surname	SDB Client information
Alias given name	TRIPS name history information
Alias surname	TRIPS name history information
Sex	SDB Client information
Date of birth	SDB Client information
Current Meshblock	TRIPS address history information
Current Statistical Area 1	TRIPS address history information
Current Statistical Area 2	TRIPS address history information
Current Statistical Area 4	TRIPS address history information
Current State	TRIPS address history information
Previous Meshblock	TRIPS address history information
Previous Statistical Area 1	TRIPS address history information
Previous Statistical Area 2	TRIPS address history information
Previous Statistical Area 4	TRIPS address history information
Previous State	TRIPS address history information
Analysis dataset	
Age at beginning of reference period	SDB Client information
Applicant status on visa application	SDB Client information
Citizenship country (Other than Australia)	SDB Client information
Country of birth	SDB Client information
Ethnicity	SDB Client information
Location of visa application	SDB Client information
Marital status	SDB Client information
Occupation on visa application	SDB Client information
Preferred language	SDB Client information
Proficiency in spoken English	SDB Client information
Relationship to offshore primary applicant	SDB Client information
Sex	SDB Client information
Religion	SDB Client information
Visa subclass	SDB Client information
Whether received Australian citizenship before beginning of reference period	SDB Client information
Year of arrival	SDB Client information

B.2 Data items available from the 2011–12 PIT linking and analysis datasets

Variable name	*Source*
Linking dataset	
Given name	Name and address register
Surname	Name and address register
Sex	Name and address register
Date of birth	Name and address register
Meshblock	Name and address register
Statistical Area 1	Name and address register
Statistical Area 2	Name and address register
Statistical Area 4	Name and address register
State	Name and address register
Analysis dataset	
Person level	
Assessable foreign source income	Client dataset
Employee income	Client dataset
HELP repayment amount	Client dataset
Foreign income	Client dataset
Industry of own unincorporated business (ANZSIC06)	Business Income Tax dataset
Investment income	Client dataset
Level of private health cover	Client dataset
Local Government Area	Client dataset
Maximum jobs held simultaneously on Job file	Individual Pay As You Go
Net rent	Client dataset
Occupation in main job (ANZSCO06)	Client dataset
Other income	Client dataset
Own unincorporated business income	Client dataset
Remoteness Area	Client dataset
Statistical Area 2	Client dataset
Superannuation and annuity income	Client dataset
Tax withheld	Client dataset
Taxable income	Client dataset
Total deductions	Client dataset
Total income	Client dataset
Total jobs held on Job file	Individual Pay As You Go
Job level	
Days in job	Individual Pay As You Go
Economic Sector of Employer (SISCA08)	Australian Business Register
Gross payments amount	Individual Pay As You Go
Industry of Employer (ANZSIC06)	Australian Business Register
Average pay per day	Individual Pay As You Go
Tax withheld	Individual Pay As You Go

C. LINKING PASSES

C.1 2016 Linking strategy for 2011–12 PITMID

The following terms are used when describing the linking strategy for the 2011–12 PITMID.

Administrative dates

Two dates appeared on the SDB with greater frequency than any other. These dates were:

- 1 January
- 31 December

These dates were used when a migrant's day and month of birth was unknown and are referred to as "administrative dates". Several passes in the linking strategy required that the value for "Date of birth" not be one of these dates in order to be considered a match.

Damerau-Levenshtein distance

The Damerau-Levenshtein distance is the distance between two strings of characters, given by counting the minimum number of operations needed to transform one string into another. Operations are defined as an insertion, deletion or substitution of a single character or the transposition of two adjacent characters.

The Damerau-Levenshtein distance differs from the Levenshtein distance in that it allows transpositions.

Several passes applied the Damerau-Levenshtein distance to first name and surname to allow for misspellings.

Levenshtein distance

The Levenshtein distance is the distance between two strings of characters, given by counting the minimum number of operations needed to transform one string into another. Operations are defined as an insertion, deletion or substitution of a single character.

The Levenshtein distance differs from the Damerau-Levenshtein distance in that it does not allow transpositions.

Several passes applied the Levenshtein distance to date of birth to allow for minor errors.

Winkler score

The Jaro-Winkler distance is a measure of similarity between two strings of characters. The higher the Jaro-Winkler distance for two strings is, the more similar the strings are. The score is normalised such that 0 equates to no similarity and 1 is an exact match.

Using the available variables, the following linking runs were conducted:

1. Run 1
   a. First name (Exact), Surname (Exact), Sex (Exact) and Date of birth (Exact) at MB and SA1;

2. Run 2
   a. First name (Exact), Surname (Exact), Sex (Exact) and Date of birth (Exact) at SA2;
   b. First name (Damerau-Levenshtein distance of 1), Surname (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   c. First name (Exact), Surname (Damerau-Levenshtein distance of 1), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   d. First name (Winkler distance of 0.93), Surname (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   e. First name (Exact), Surname (Winkler distance of 0.93), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;

3. Run 3
   a. First name (Winkler distance of 0.91), Surname (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   b. First name (Exact), Surname (Winkler distance of 0.91), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   c. First name (Winkler distance of 0.91), Surname (Winkler distance of 0.91), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2
   d. All names in alphabetical order (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   e. All names in alphabetical order (Damerau-Levenshtein distance of 2), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   f. Alias first name (Exact), Alias surname (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   g. First initial (Exact), Surname (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
   h. First name (Exact), Surname (Exact), Sex (Exact) and Date of birth (no administrative dates) at MB, SA1 and SA2;
   i. First name (Exact), Surname (Exact), Sex (Exact) and Date of birth (Levenshtein distance of 1) at MB, SA1 and SA2;

4. Run 4
    a. First name (Exact), Surname (Exact), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    b. First name (Damerau-Levenshtein distance of 1), Surname (Exact), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    c. First name (Exact), Surname (Damerau-Levenshtein distance of 1), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    d. First name (Winkler distance of 0.93, 0.91 and 0.85), Surname (Exact), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    e. First name (Exact), Surname (Winkler distance of 0.93, 0.91 and 0.85), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    f. All names in alphabetical order (Exact), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia
    g. All names in alphabetical order (Damerau-Levenshtein distance of 2), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    h. Alias first name (Exact), Alias surname (Exact), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    i. First initial (Exact), Surname (Exact), Sex (Exact) and Date of birth (Exact) at SA4, State and Australia;
    j. First name (Exact), Surname (Exact), Sex (Exact) and Date of birth (no administrative dates) at SA4, State and Australia
    k. First name (Exact), Surname (Exact), Sex (Exact) and Date of birth (Levenshtein distance of 1) at SA4, State and Australia;

5. Run 5
    a. Single first name (Exact), Single surname (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;

6. Run 6
    a. Surname (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
    b. First name (Exact), Sex (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
    c. First name (Exact), Surname (Exact) and Date of birth (Exact) at MB, SA1 and SA2;
    d. First name (Exact), Surname (Exact) and Sex (Exact) at MB, SA1 and SA2;

C.1 2016 Linking runs and results for 2011–12 PITMID

	First name	*Surname*	*All names*	*Alias first name*	*Alias surname*	*First initial*	*Sex*	*Date of birth*	*Geography*	*No. Links generated*
Run 1										
	E	E					E	E	MB	12,170
	E	E					E	E	SA1	37,636
Run 2										
	E	E					E	E	SA2	17,023
	DL1	E					E	E	MB	40
	DL1	E					E	E	SA1	11
	DL1	E					E	E	SA2	9,361
	E	DL1					E	E	MB	18
	E	DL1					E	E	SA1	8
	E	DL1					E	E	SA2	115
	W93	E					E	E	MB	15
	W93	E					E	E	SA1	1
	W93	E					E	E	SA2	183
	E	W93					E	E	MB	6
	E	W93					E	E	SA1	0
	E	W93					E	E	SA2	521
Run 3										
	W91	E					E	E	MB	0
	W91	E					E	E	SA1	0
	W91	E					E	E	SA2	7
	E	W91					E	E	MB	0
	E	W91					E	E	SA1	0
	E	W91					E	E	SA2	0
	W91	W91					E	E	MB	0
	W91	W91					E	E	SA1	12
	W91	W91					E	E	SA2	7
			E				E	E	MB	0
			E				E	E	SA1	0
			E				E	E	SA2	577
			DL2				E	E	MB	0
			DL2				E	E	SA1	0
			DL2				E	E	SA2	624
				E	E		E	E	MB	0
				E	E		E	E	SA1	0
				E	E		E	E	SA2	2,754
		E				E	E	E	MB	416
		E				E	E	E	SA1	878
		E				E	E	E	SA2	776
	E	E					E	AD E	MB	0
	E	E					E	AD E	SA1	0
	E	E					E	AD E	SA2	0
	E	E					E	L1	MB	195
	E	E					E	L1	SA1	758
	E	E					E	L1	SA2	451
Run 4										
	E	E					E	E	SA4	10,626
	E	E					E	E	State	0
	E	E					E	E	Australia	26,215
	DL1	E					E	E	SA4	7
	DL1	E					E	E	State	0
	DL1	E					E	E	Australia	182
	E	DL1					E	E	SA4	22
	E	DL1					E	E	State	0
	E	DL1					E	E	Australia	125
	W93	E					E	E	SA4	3,267
	W93	E					E	E	State	0
	W93	E					E	E	Australia	3,974
	E	W93					E	E	SA4	0
	E	W93					E	E	State	0
	E	W93					E	E	Australia	149
	W91	E					E	E	SA4	0
	W91	E					E	E	State	0
	W91	E					E	E	Australia	242

C.1 2016 Linking runs and results for 2011–12 PITMID — continued

	First name	*Surname*	*All names*	*Alias first name*	*Alias surname*	*First initial*	*Sex*	*Date of birth*	*Geography*	*No. Links generated*
Run 4 (cont.)										
	E	W91					E	E	SA4	12
	E	W91					E	E	State	0
	E	W91					E	E	Australia	43
	W85	E					E	E	SA4	349
	W85	E					E	E	State	0
	W85	E					E	E	Australia	322
	E	W85					E	E	SA4	119
	E	W85					E	E	State	0
	E	W85					E	E	Australia	197
			E				E	E	SA4	523
			E				E	E	State	0
			E				E	E	Australia	1,530
			DL2				E	E	SA4	41,079
			DL2				E	E	State	0
			DL2				E	E	Australia	1,346
				E	E		E	E	SA4	1,403
				E	E		E	E	State	0
				E	E		E	E	Australia	3,553
		E				E	E	E	SA4	860
		E				E	E	E	State	0
		E				E	E	E	Australia	4,306
	E	E					E	AD E	SA4	0
	E	E					E	AD E	State	0
	E	E					E	AD E	Australia	0
	E	E					E	L1	SA4	1,462
	E	E					E	L1	State	0
	E	E					E	L1	Australia	8,924
Run 5										
	S E	S E					E	E	MB	0
	S E	S E					E	E	SA1	0
	S E	S E					E	E	SA2	1
Run 6										
	E						E	E	MB	394
	E						E	E	SA1	2,319
	E						E	E	SA2	2,896
		E					E	E	MB	165
		E					E	E	SA1	246
		E					E	E	SA2	1,900
	E	E						E	MB	109
	E	E						E	SA1	184
	E	E						E	SA2	411
	E	E					E		MB	399
	E	E					E		SA1	1,613
	E	E					E		SA2	6,610
Total										212,647

AD	No Administrative dates
DL1	Damerau-Levenshtein distance of 1
DL2	Damerau-Levenshtein distance of 2
E	Exact match
L1	Levenshtein distance of 1
MB	Meshblock
S	Single name
SA1	Statistical Area 1
SA2	Statistical Area 2
W85	Winkler score of 0.85
W91	Winkler score of 0.91
W93	Winkler score of 0.93

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