



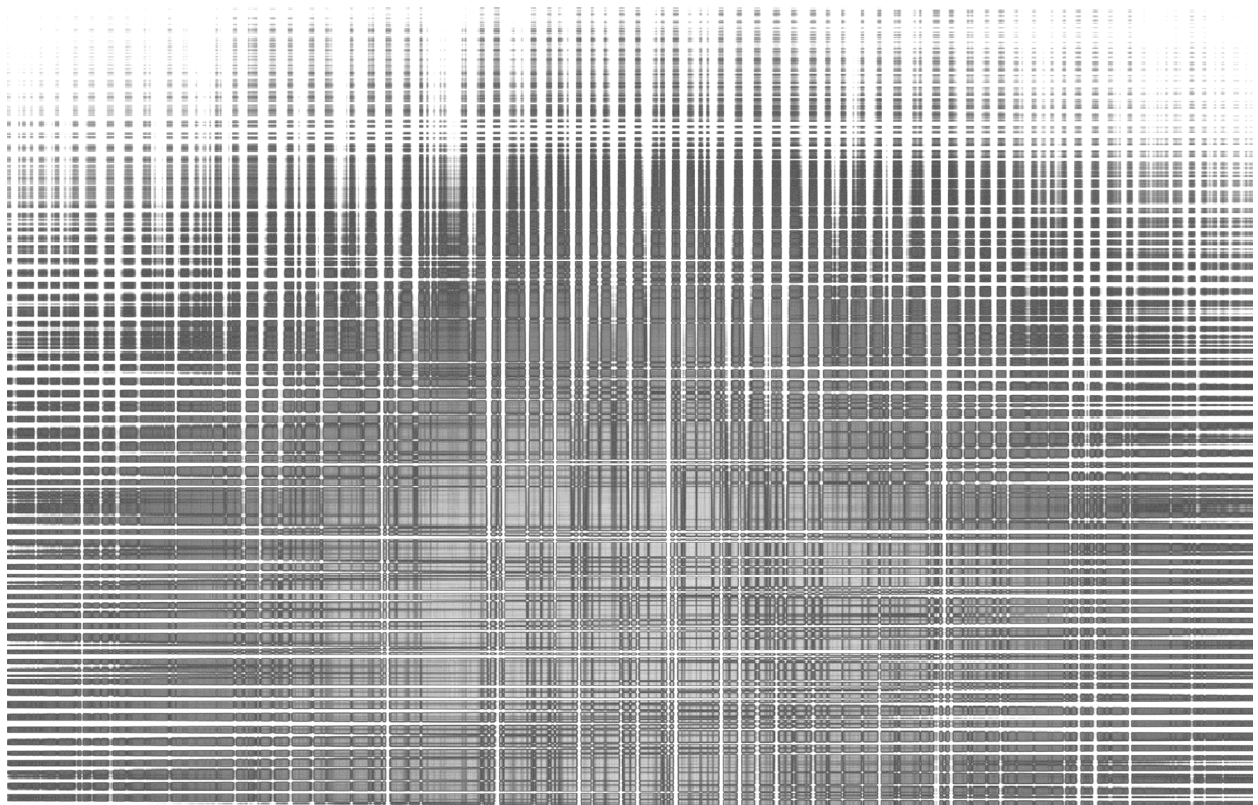
Australian  
National  
University



Social  
Research  
Centre

An ANU Enterprise business

CENTRE FOR SOCIAL  
RESEARCH & METHODS



# Introducing the longitudinal MADIP and its role in understanding income dynamics in Australia

**N Biddle, R Breunig, F Markham and C Wokker**

CSRM & SRC METHODS PAPER

NO. 1/2019

---

## Series note

The ANU Centre for Social Research & Methods (CSRM) was established in 2015 to provide national leadership in the study of Australian society. CSRM has a strategic focus on:

- development of social research methods
- analysis of social issues and policy
- training in social science methods
- providing access to social scientific data.

CSRM publications are produced to enable widespread discussion and comment, and are available for free download from the CSRM website (<http://csmr.cass.anu.edu.au/research/publications>).

CSRM is located within the Research School of Social Sciences in the College of Arts & Social Sciences at the Australian National University (ANU). The centre is a joint initiative between

the Social Research Centre and the ANU. Its expertise includes quantitative, qualitative and experimental research methodologies; public opinion and behaviour measurement; survey design; data collection and analysis; data archiving and management; and professional education in social research methods.

As with all CSRM publications, the views expressed in this Methods Paper are those of the authors and do not reflect any official CSRM position.

### **Professor Matthew Gray**

Director, ANU Centre for Social Research & Methods  
Research School of Social Sciences  
College of Arts & Social Sciences  
The Australian National University  
March 2019

Methods Paper No. 1/2019

ISSN 2209-184X

ISBN 978-1-925715-18-7

An electronic publication downloaded from  
<http://csmr.cass.anu.edu.au/research/publications>.

For a complete list of CSRM methods papers, see  
<http://csmr.cass.anu.edu.au/research/publications/methods-research-papers>.

### **ANU Centre for Social Research & Methods**

Research School of Social Sciences  
The Australian National University

---

# Introducing the longitudinal MADIP and its role in understanding income dynamics in Australia

**N Biddle, R Breunig, F Markham and C Wokker**

**Nicholas Biddle** is the Associate Director of the ANU Centre for Social Research & Methods, and a Fellow of the Tax and Transfer Policy Institute.

**Robert Breunig** is Director of the Tax and Transfer Policy Institute.

**Francis Markham** is a Research Fellow at the ANU Centre for Aboriginal Economic Policy Research.

**Chris Wokker** is a PhD student at the Tax and Transfer Policy Institute.

---

## Abstract

Understanding the determinants, dynamics and distribution of income within a country is an area of ongoing research and policy interest. There is a lot we do know about income dynamics in Australia. However, we have limited information on several key aspects, including detailed information on the spatial dimensions of income inequality, income mobility in and out of the top of the distribution, sociodemographic characteristics at the top of the distribution, the top of the household income distribution, the impact of local geographic inequality and other characteristics on income change, and the complete income distribution of small subpopulations of particular policy interest. Our knowledge has been limited by a lack of data germane to these issues. The opening up of access to data from the Multi-Agency Data Integration Project (MADIP),

including the Basic Longitudinal Extract 2011 (BLE2011), provides an opportunity to fill some of these research gaps. The aim of this paper is to begin the external validation of these data, with a particular focus on what the data can tell us about the distribution, dynamics and determinants of income in Australia. Our view is that the BLE2011 has the potential to shed new light on these aspects. However, analysis of the dataset should be done with caution, taking into account some key limitations, including incomplete linkage (which is likely to be nonrandom) and more limited income information for those who do not complete a tax return. These limitations notwithstanding, the BLE2011 and other datasets from MADIP should form an important part of the social science infrastructure in Australia.

---

## Acknowledgments

The data for this paper were provided through the Australian Bureau of Statistics (ABS) Virtual Data Laboratory. The authors would like to thank the ABS for the collegial approach to data access and recognise the considerable efforts to make data available in a way that maintains data privacy. The authors would also like to thank two reviewers from within the ANU Centre for Social Research & Methods: Dr Rob Bray and Associate Professor Ben Phillips.

---

## Acronyms

ABS	Australian Bureau of Statistics
ANU	Australian National University
BLE2011	Basic Longitudinal Extract 2011
CSRM	ANU Centre for Social Research & Methods
HILDA	Household, Income and Labour Dynamics in Australia
MADIP	Multi-Agency Data Integration Project
OECD	Organisation for Economic Co-operation and Development
PIT	personal income tax
SSRI	Social Security and Related Information

---

# Contents

Abstract	iii
Acknowledgments	iv
Acronyms	iv
1. Income and its determinants in Australia	1
2. Data and data gaps in Australia	2
2.1 Personal income tax records	2
2.2 Longitudinal household surveys	2
2.3 Cross-sectional household surveys	3
2.4 Census of Population and Housing	3
2.5 Unknowns regarding income distribution in Australia	3
3. The Multi-Agency Data Integration Project	5
4. MADIP and other Australian data on income	7
5. Income data in MADIP	9
5.1 For how many people do we have PIT records in the BLE2011, cross-sectionally and longitudinally?	9
5.2 How does income in the tax database compare with income in the census?	11
5.3 What are the predictors of variation between census and PIT income data?	14
5.4 What are the characteristics of those with missing census data in the BLE2011?	17
6. Concluding comments and future analysis using MADIP	19
Notes	20
References	20

---

## Tables and figures

Table 1	Datasets and years available in the MADIP Basic Longitudinal Extract 2011	6
Table 2	Characteristics of major datasets on income in Australia	7
Figure 1	Number of years with tax data for those in BLE2011	10
Table 3	Factors associated with number of years of personal income tax data, linked census records, people aged 15 years or over in 2011	10
Figure 2	Relationship between census income (2011) and PIT income (2010–11 financial year) in nominal dollars, adults aged 15 years and over in 2011	12
Figure 3	Percentage of people with not-stated income in the census across the PIT income distribution (2010–11), adults aged 15 years and over in 2011	14

Figure 4	Percentage of sample whose PIT income is below, within or above their census income category	15
Table 4	Factors associated with whether PIT income is below, within or above census income category	16
Table 5	Demographic and socioeconomic information by census linkage status and for the total census population	18



# 1. Income and its determinants in Australia

The determinants of the distribution of income in Australia differ substantially from those in other comparable countries. Australia's social security system targets benefits to the relatively disadvantaged to a much greater extent than most other countries in the Organisation for Economic Co-operation and Development (OECD) (Whiteford 2010). Australia was much less affected by the Great Recession than the systems in other countries (Claessens et al. 2010) and has a number of population groups that receive targeted policy attention, including the Aboriginal and Torres Strait Islander population (Markham & Biddle 2018). Geographically, Australia is a large, sparsely populated continent, containing 5.9% of the world's land but only 0.3% of the world's population (World Bank 2018). It is highly urbanised, with an urban structure that concentrates the national population in its two largest cities to an unusual degree (Ellis & Andrews 2001). All of this means that the determinants of income in Australia and recent dynamics are likely to differ from those of other countries, as is any policy response to distributional questions.

Australia's distinctiveness does not mean that it has been exempt from the increasing attention paid to income distribution within (and across) countries. Indeed, the Productivity Commission (2018) recently undertook a 'stocktake of the evidence'; with regard to income change over the past 2–3 decades, it concluded that:

*what ... distinguishes Australia from most other developed countries has been its unprecedented 27-year period of uninterrupted economic growth ... it has delivered for the average Australian household in every income decile significantly improved living standards.*

Despite this reasonably positive finding, the Productivity Commission also showed that income growth has been very different across the age distribution, with the income of the young growing relatively slowly in recent years (compared with growth in income of the relatively young for previous cohorts). However, in Australia, unlike in some other developed countries, there has still been income growth across generations. We have shown in a few different contexts that income growth has not been consistent or equally shared in different geographical regions (Biddle & Montaigne 2017) or with regard to Indigenous status (Markham & Biddle 2018). These issues were not a focus of the Productivity Commission's analysis. In aggregate, however, recent work by Wilkins (2015) suggests that income inequality has been relatively stable during the past two decades, notwithstanding the concurrent increase in the share of income accruing to the top 1% of the population over this period (Burkhauser et al. 2015).

There has also been significant debate in Australia about the adequacy of certain forms of transfer payments. In particular, payments for jobseekers (Newstart Allowance) are indexed to inflation, whereas the age pension is indexed to earnings (or inflation, if that is higher). This means that incomes for these two groups have diverged; those who receive the age pension, but not jobseekers, have experienced improvements in living standards from productivity increases (OECD 2010).



## 2. Data and data gaps in Australia

Although we have a reasonable understanding of income dynamics in Australia, with an active academic research community in this area (Wilkins 2015), there is much that we do not know. This information gap has been driven not by a lack of interest but primarily by a lack of data. Partly, this is because the policy settings in Australia have not required the gathering of such data. Australia's social insurance system is not based on wages, and hence there are no complete administrative data on wages and salaries. Furthermore, we do not have any form of national register, which other countries often use to link administrative data to. However, there have also been barriers to accessing the data that are available.

The four main sources of income data used in Australia for research on inequality and income dynamics are described below.

### 2.1 Personal income tax records

Personal income tax (PIT) records are administrative data covering all taxpayers. The data cover people at the top of the income distribution, and are therefore most prominently used for studies of top income shares (e.g. Leigh 2005, Atkinson & Leigh 2007, Gaston & Rajaguru 2009, Burkhauser et al. 2015). However, taxation data that have been made available to researchers to date have been cross-sectional only. They provide limited information about wealth, geography or sociodemographic variables. Because Australia's income tax unit is the individual, PIT data only contain limited information at the household level. For many records, it is possible to link income for spouses. However, not all spouses need to report income, and there is no information on other members of the household.

Finally, PIT records only contain data on taxable income (i.e. they exclude income from tax-exempt sources such as scholarships). They do not include information on those who do not complete a tax return in a given year – a group of around 40% of adults who are typically at

the bottom of the personal income distribution (Atkinson & Leigh 2005).

### 2.2 Longitudinal household surveys

One of the richest datasets in Australia for understanding income and its distribution is the Household, Income and Labour Dynamics in Australia (HILDA) survey. HILDA contains extensive data on income, wealth and consumption for a longitudinal panel of Australian households (9631 households in wave 15). Crucially, HILDA's longitudinal design enables the study of income dynamics, including mobility within the income distribution. HILDA also contains rich detail on sociodemographic variables and data on all household members, making possible analyses of income inequality that take households as the economic unit of analysis (Saunders & Bradbury 2006; Finlay 2012; Wilkins 2014a, 2015; Burkhauser et al. 2015; Productivity Commission 2018).

However, because of its relatively small sample size, HILDA may miss the dynamics of small groups, such as at the very top of the income distribution. In addition, across the income distribution, there are limits in using a sample survey to analyse income dynamics, because measurement error across multiple waves can appear as instability in income (Solon 1989).

Like any longitudinal survey, HILDA is prone to differential nonresponse and, more importantly, nonrandom attrition. Those who drop out of HILDA are likely to be very different from those who stay. Despite considerable efforts, HILDA has become less representative through time. Sample attrition has been highest among those born in a non-English-speaking country, the unemployed, those working in low-skilled occupations and the young (Summerfield et al. 2017), potentially biasing analyses of population-level trends in income inequality if these trends are different for different groups.

## 2.3 Cross-sectional household surveys

The official data on income distribution in Australia come from the long-running Survey of Income and Housing (SIH), currently conducted every 2 years by the Australian Bureau of Statistics (ABS), and its add-on component, the Household Expenditure Survey, conducted every 6 years. The predecessor to the SIH was known as the Income Distribution Survey. The SIH is arguably more representative than HILDA (at least in later waves of HILDA) and may have the highest-quality data on income and wealth among people sampled; it is therefore the dataset that is most often used to analyse income inequality in Australia (Saunders 1993, 2004; Harding 1995, 1997; Barrett et al. 2000; Athanasopoulos & Vahid 2003; Johnson & Wilkins 2004; Saunders & Bradbury 2006; Productivity Commission 2018). Furthermore, the frequency of data collection has increased and methodological changes have become less common over recent years.

However, the SIH is cross-sectional, and the previous methodological changes over time mean that comparisons of repeated cross-sections over a longer period can be potentially misleading (Wilkins 2014a). Because of the SIH's modest sample size, it has a relatively small sample of key points on the income distribution. The SIH's modest sample size, and focus on income and wealth also mean that limited information is available on small groups of particular interest to policy makers (e.g. Indigenous Australians, certain migrant cohorts, single-parent families, people with a disability).

## 2.4 Census of Population and Housing

The Australian Census of Population and Housing is conducted every 5 years. Unlike in some other countries, it is a full, compulsory, long-form census. Although it includes a question on income, item response rates can be very low (91.0% in 2016). As well, information on income is collected in ranges that have tended to be quite coarse, and change over time as a result of both inflation and question redesign. For the 2016 Census, the top income category contained 3.4%

of individuals aged 15 years or older who stated their income.

Unlike the other income sources mentioned above, particularly the two surveys, the census does not have good information on wealth, nor does it have information on after-tax or disposable income. Although the income measures in the census are highly correlated with these more detailed measures of access to economic resources, it is often the differences between the measures that are of interest.

Despite these limitations, the census is the only data source to adequately cover small geographic areas and small subpopulations, and for that reason remains a key data source for understanding the distribution of income in Australia (e.g. Hunter & Gregory 1996, Athanasopoulos & Vahid 2003, Hunter 2003, Leigh 2006, Fleming & Measham 2015, Biddle & Montaigne 2017, Markham & Biddle 2018). Furthermore, the recently created linked Australian Census Longitudinal Database (Chipperfield et al. 2017) provides a rudimentary, but very valuable, tool for understanding income dynamics and mobility for these and other groups.

## 2.5 Unknowns regarding income distribution in Australia

Given the limitations of income data in Australia outlined above, there is much about the distribution of income and income dynamics in Australia that we do not know. First, there are limits to the available information on the spatial dimensions of income inequality, particularly in very small geographic areas such as neighbourhoods and suburbs. Only census data provide sufficient geographical coverage to examine the entire population, but income data are collected in quite coarse ranges, are subject to substantial top-coding and only report total gross income (without differentiating income by source or disposable income). A variety of approaches have been used to address these difficulties, but they are all associated with limitations (Hunter & Gregory 1996, Hunter 2003, Leigh 2006, Biddle & Montaigne 2017).

We know very little about **income mobility, particularly in and out of the top of the distribution**. Several studies have used HILDA to examine mobility in and out of the top income quintile (Rohde et al. 2010; Wilkins 2014b, 2016; Kaplan et al. 2018), and poverty dynamics have been subject to more scrutiny (e.g. Headey et al. 2005, Abello & Harding 2006, Saunders & Bradbury 2006, Buddelmeyer & Verick 2008, Wilkins 2016, Venn & Hunter 2018). Little is known, however, about mobility for those receiving top incomes because the tax data that researchers have access to are cross-sectional. As well, longitudinal surveys only have a small number of people at the very top of the income distribution, and are prone to significant sampling and nonsampling error.

Furthermore, we do not have a rich picture of the **sociodemographic characteristics** of Australians with very high incomes. Tax data only include the information that is required for people to complete their tax obligations; the survey data are subject to sampling error and nonsampling error; and data collections exclusively devoted to this subject are unlikely to be representative of the relevant part of the Australian population (e.g. Wiesel 2018).

Recently, the Australian Tax Office released the Australian Longitudinal Individuals File (Alife).<sup>1</sup> This is a 10%, longitudinal sample of individual taxpayers that is available for approved researchers through remote data access. It was not available in time for validation in this paper, but may be useful in future as data for comparison with the Multi-Agency Data Integration Project (MADIP) to look at movement in and out of the top of the distribution. Limitations of Alife compared with the BLE2011 is that it has limited demographic information, no census-based income and no information on the social security system.

Although some household information must be entered on tax returns, we do not have as much information on the top of the **household income distribution** as we do on the top of the individual income distribution. Assortative mating, differential household size and sharing of income (Greenwood et al. 2014) mean that the income distribution for households may be quite different from the income distribution for individuals.

We do not know whether the individuals at the top or bottom of the individual income distribution are the same as those in the households at the top or bottom of the household income distribution.

Australia has three levels of government (national, state/territory and local), and policy varies at each of these levels. Furthermore, an increasing amount of research shows that a person's income is influenced by complex agglomeration effects (Beeson 2017), and that the characteristics of those who live in close geographic proximity to an individual potentially, but not necessarily, affect that individual's outcomes. The lack of longitudinal data and information on the bottom of the income distribution means that our information on the **impact on income change of local geographic inequality and other characteristics** is limited (e.g. Leigh 2006); surveys do not have large enough samples at that scale.

Finally, Australia has a number of **small subpopulations** that are of particular interest to policy makers for whom we have only limited income information. For example, a number of specific policy initiatives and targets apply to the Aboriginal and Torres Strait Islander population (Biddle et al. 2017). However, because Indigenous Australians make up only 3.3% of the population (and a disproportionate percentage of Indigenous Australians live in areas that are expensive to survey), limited information is available from our main income surveys for this group, and sampling errors are large. Furthermore, there is no Indigenous identifier in our tax data, and, in any case, Indigenous Australians are concentrated at the lower part of the income distribution where tax data are less comprehensive. Although some Indigenous-specific surveys have income information (e.g. the National Aboriginal and Torres Strait Islander Social Survey), these surveys are infrequent and do not have a non-Indigenous comparison. Even less information is available about other population groups of similar policy interest (e.g. recent migrant cohorts and refugees, people with disability).

### 3. The Multi-Agency Data Integration Project

Given the gaps in data and knowledge outlined above, we need to develop and make use of new datasets in Australia that have longitudinal information across the income distribution, for a large sample of individuals, that include sociodemographic and household information linked to individual records.

A recently developed dataset that meets these criteria is the MADIP Basic Longitudinal Extract 2011 (BLE2011). According to the summary provided by the ABS,<sup>2</sup> the BLE2011 is built around a full cohort of the Australian population in 2011 and:

*... includes key demographic, social, health care, government payment and income information for this population over the period 2011–2016. The microdata product contains approximately 22.5 million records and 122 data items; 74 of these data items have information for multiple years to enable longitudinal analysis.*

Four sources of data in the BLE2011 have been linked (probabilistically) at the individual level:

- the Medicare Enrolments Database and Medicare Benefits Schedule data, which include information on the number of services used, benefits paid and fees charged
- PIT data, which include information on wages and salaries; government allowances, pensions and payments; total income (summation of the previous, as well as other forms of income); and taxable income
- Social Security and Related Information (SSRI) data, which include information on whether a person was receiving income from 28 separate payments at a September snapshot, and whether their partner was receiving any of the payments
- a subset of data from the 2011 Census, which includes a household identifier; and information on ethnicity, country of birth, education levels and participation,

employment status (including hours worked, industry and occupation), disability status and income.

In addition to data from the individual datasets, a small number of derived variables combine information from multiple datasets. These are age, sex, Indigenous status and geographic location as of 2011. The latter is available at the level of Statistical Area Level 1, a very detailed level of geography.<sup>3</sup>

Table 1 summarises the years and data available in the BLE2011.

The datasets were linked using different methods, depending on the data, although none were linked using a unique identifier. The first step in the linkage was to (separately) link valid PIT data to Medicare and valid SSRI data to Medicare. Both sets of linkage were undertaken using anonymised name, sex, geo-coded address and date of birth. High-quality linkages were prioritised, and 93.4% of PIT data and 94.4% of SSRI records were linked to a Medicare record.

The second step was to link census data. The data were linked to Medicare and SSRI data using slightly different information. Because of a commitment by the ABS to not use names from the 2011 Census for data linkage purposes, much less information could be used, and linkage rates were therefore lower. For the census to Medicare linkage, the ABS used sex, geo-coded residential address, date of birth/age and number of persons in a household. For the census to SSRI linkage, the ABS used (in addition) partner's date of birth, country of birth, Indigenous status, marital status and children's dates of birth. The linkage rate achieved was 66.5%.

The final step in creating the BLE2011 was to create longitudinal information for the Medicare, PIT and SSRI data. Although it is not completely clear from the documentation, it would appear that the longitudinal data are then linked within

**Table 1 Datasets and years available in the MADIP Basic Longitudinal Extract 2011**

Year	Census	Medicare	PIT	SSRI
2010	na	na	2010–11 financial year	na
2011	2011 cross-sectional	2011 calendar year services and benefits	2011–12 financial year	September 2011 payment type
2012	na	2012 calendar year services and benefits	2012–13 financial year	September 2012 payment type
2013	na	2013 calendar year services and benefits	2013–14 financial year	September 2013 payment type
2014	na	2014 calendar year services and benefits	2014–15 financial year	September 2014 payment type
2015	na	2015 calendar year services and benefits	2015–16 financial year	September 2015 payment type
2016	na	2016 calendar year services and benefits	na	September 2016 payment type

MADIP = Multi-Agency Data Integration Project; na = not applicable; PIT = personal income tax; SSRI = Social Security and Related Information

each dataset using a unique identifier from within those datasets. That is, the 2011–12 PIT data are linked to the 2010–11 PIT data using tax file number; the 2012 SSRI data are linked to the 2011 SSRI data using customer reference number; and the 2012 Medicare data are linked to the 2011 Medicare data using Medicare number.

It is possible, and perfectly feasible, to create longitudinal census information for the BLE2011. This could be done by linking the 2011 Census to the 2016 Census directly (as has been done with the Australian Census Longitudinal Dataset) or

by linking the 2016 Census to the 2016 Medicare and/or SSRI data. However, to maintain a commitment made by the ABS to link only 5% of the 2011 Census data, this has not occurred to date (and may not occur).

Although linkage was incomplete across the four datasets, the data available to researchers (BLE2011) have both linked and unlinked individuals. There are, of course, individuals who are otherwise in scope but missing from all four datasets.

## 4. MADIP and other Australian data on income

If income distribution and dynamics were a landscape, each of the datasets surveyed here might be said to provide a different window through which to view that landscape – each angle of view reveals new features that were hitherto obscured. Table 2 summarises the similarities and differences between the MADIP

BLE2011 and other data sources for studying income in Australia. It demonstrates that the MADIP BLE2011 has a combination of unique characteristics not available in other Australian data sources, filling some of the data gaps identified earlier in the paper.

**Table 2** Characteristics of major datasets on income in Australia

Characteristic	PIT	SIH/HES	HILDA	Census	MADIP BLE2011
Collection type	Administrative collection	Survey	Survey	Census	Census linked to three administrative collections
Design	Repeated cross-sectional	Repeated cross-sectional	Panel (currently 17 waves)	Repeated cross-sectional	Cross-sectional census linked to administrative panels
Scope	People who submit a tax return	Usual residents of private dwellings, excluding very remote areas	Usual residents of private dwellings, excluding very remote areas	People who spent Census night in Australia	All Australian residents in 2011
Income units	Persons	Persons, income units and households	Persons, income units and households	Persons and households	Persons (all years) and households (census year)
Frequency	Annual	Currently every 2 years; previously either every 3 years or annual	Annual	Every 5 years	Every 5 years, with annual administrative panels
First collected	1921	1979	2001	1901 (income included since 1976)	2011
Sample size (most recent collection)	13.5 million	33 968 people in 17 768 households	17 571 people in 9742 households	23.8 million people in 9.9 million dwellings	22.5 million people
Sources of income	Disaggregated by source (excluding tax-exempt sources)	Detailed indicators of income sources	Detailed indicators of income sources	No indicators of income sources	Reports total income, taxable income (income after deductions), wages/salaries and social security income

**Table 2** *continued*

Characteristic	PIT	SIH/HES	HILDA	Census	MADIP BLE2011
Gross or disposable income	Gross	Both	Both	Gross	Gross, but with approximations of disposable feasible
Censoring income variables (most recent collection)	Aggregate tables	None	None	Income coded to ranges; top-coded at \$156 000	Income measures top-coded above \$250 000
Geographical detail	Typically national, with limited geographic tabulations available	Greater capital city Statistical Areas	Coded to small areas (SA1s), but with very few respondents in each area	Coded to very small areas (mesh blocks), enumerating all residents in each area	Coded to very small areas (mesh blocks), enumerating all residents in each area
Sociodemographic variables	Few	Some	Many	Some	Some

HES = Household Expenditure Survey; HILDA = Household, Income and Labour Dynamics in Australia; MADIP BLE2011 = Multi-Agency Data Integration Project Basic Longitudinal Extract 2011; PIT = personal income tax; SA1 = Statistical Area Level 1; SIH = Survey of Income and Housing

## 5. Income data in MADIP

The BLE2011 is a unique dataset that allows us to draw new insights about the determinants, dynamics and distribution of income in Australia. It fills the gap outlined earlier by providing longitudinal income information across the distribution, for a large sample of individuals, with sociodemographic and household information at the individual level.

Despite the potential benefits of the BLE2011 for understanding income in Australia, there are a number of limitations to the dataset and associated questions regarding its usability. Some of these limitations are shared by the component datasets, as well as by other datasets that researchers have used in the past to understand income and its distribution. However, other limitations – such as the relatively low linkage rate with the census – are unique to this type of dataset.

In this section, we consider four data validation questions, with a particular focus on the tax data. The answers to these questions will strongly influence the use of the data for inequality research. The questions are as follows:

- For how many people do we have PIT records in the BLE2011, cross-sectionally and longitudinally?
- How does income in the tax database compare with income in the census?
- What are the predictors of variation between census and PIT income data?
- What are the characteristics of those with missing census data in the BLE2011 (either because the individuals did not complete the census or the census data could not be linked)?

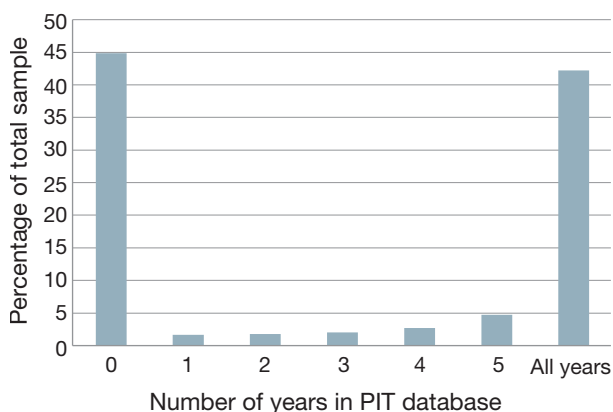
### 5.1 For how many people do we have PIT records in the BLE2011, cross-sectionally and longitudinally?

Despite the claim on the website of the Australian Tax Office that ‘most people need to lodge a tax return each year’, the reality is that a very large proportion of people in Australia do not. Furthermore, many individuals complete their tax returns late. One group of individuals who do not need to lodge a tax return are those whose taxable income is below the tax-free threshold (and hence have a zero marginal and average rate of tax). Counterintuitively, this group includes many people whose income exceeds the tax-free threshold, but whose income contains a substantial tax-exempt component (e.g. certain social security payments and scholarships, most inheritances, and some forms of superannuation income). This means that not everyone in Australia for whom we would like to have income information is in the PIT database.

One group of individuals who do not need to lodge a tax return are those whose income is below the tax-free threshold (and hence have a zero marginal and average rate of tax). For the 2010–11 financial year (the first year of PIT data in the BLE2011), the tax-free threshold was \$6000 (in nominal terms); by the 2015–16 financial year, this had risen to \$18 200. Figure 1 gives the proportion of people in the BLE2011 by the number of years that they have a tax record.

The BLE2011 contains two main populations (based on their tax data). The first, and slightly larger, population is those who have no tax record at all. These individuals make up around 44.9% of the BLE2011 population. The second population, making up around 42.2% of the sample, have a tax record for all 6 years in the MADIP. This leaves about 12.9% of the sample who have data for 1–5 financial years.

**Figure 1** Number of years with tax data for those in BLE2011



Source: Customised data from the MADIP Basic Longitudinal Extract 2011

Not surprisingly, a large number of those who do not have a tax record at all are under the age of 15 (around 42.2%). However, when these children are removed from the analysis, there is still a very large proportion of adults (32.0%) not in the tax database.

A number of additional characteristics (from the 2011 Census) vary based on the number of years that a person has a tax record.

For the full sample of adults (aged 15 years or over at baseline), average age is higher for those who do not have a tax record (47.5 years compared with 40.4 years for those with 6 years of data), and individuals without a tax record are much more likely to be female than those who do have a tax record (54.6% compared with 48.0%).

Age and sex are two of the few variables that are available for all individuals in the BLE2011, as these questions are asked consistently for the Medicare, PIT and SSRI data. However, for individuals who have a census record, a range of other demographic and socioeconomic variables can also be used to analyse presence in the tax system. Results are given in Table 3, based on a Poisson model, with the count of the number of years in the tax record as the dependent variable, and a linear and quadratic age term.

**Table 3** Factors associated with number of years of personal income tax data, linked census records, people aged 15 years or over in 2011

Explanatory variable	Coefficient	Marginal effect
Age	0.0256*	0.0638
Age squared	-0.0003*	
Female	0.0031*	0.0076
Indigenous	-0.1838*	-0.4130
Employed	0.8434*	3.2576
Completed year 9 or less	-0.3582*	-0.7406
Completed year 10 or 11	-0.0851*	-0.2006
Completed year 12 (omitted category)		
Has a degree	0.0211*	0.0524
Born overseas	-0.0776*	-0.1838
Midpoint of census income category (\$'0,000)	0.0152*	0.0377
Constant/predicted years for base case	0.2807*	2.4599
Sample size	9 935 074	
Pseudo R-squared	0.1662	

Notes:

1. The base-case individual is aged 40, male, non-Indigenous and not employed; has completed year 12; does not have a degree; was born in Australia; and has a census income of \$30 000 at baseline.
2. Coefficients that are significantly different from 0 at the 1% level of significance are marked with \*.

Controlling for age and sex, Indigenous Australians, people with a relatively low level of education (less than year 12) and those born overseas have fewer years in the tax data. Those who were employed at baseline and those with at least a bachelor degree have significantly and substantially more years in the linked tax data. However, these correlates account for a relatively small amount of the variation in the number of years individuals are present in the PIT data, suggesting that other unobserved characteristics play a much greater role in determining whether an individual will lodge a tax return.

The final variable included in the model is income, as recorded in the census. Not surprisingly, as census income goes up, the number of years in the tax data also goes up. However, that does not mean that those with zero years of tax data always have zero income according to the census. Indeed, if people are assigned to the midpoint of the census income band that they belong to (and \$130 000 for the upper income band), the average income as of 2011 for those without any tax data is around \$16 000. Although this is substantially lower than the average census income for those in the tax system for all 6 years (around \$51 000), it is clear that many people who are not in the tax system still report an income in the census that is well above the tax-free threshold. We discuss this finding in more detail in the next section.

## 5.2 How does income in the tax database compare with income in the census?

The income measure in the census and that captured through the PIT data are conceptually quite similar. The question asked in the 2011 Census relating to income is: ‘What is the total of all wages/salaries, government benefits, pensions, allowances and other income the person usually receives?’ People are asked to ‘not deduct: tax, superannuation contributions, health insurance, amounts salary sacrificed, or any other automatic deductions’ and are given a long list of items to include under three headings:

- pensions/allowances
  - Family Tax Benefit
  - Parenting Payment

- unemployment benefits
- Newstart Allowance
- Rent Assistance
- pensions
- student allowances
- maintenance (child support)
- workers’ compensation
- any other pensions/allowances
- other income
  - interest
  - dividends
  - rents (excluding expenses of operation)
  - business/farm income (excluding expenses of operation)
  - income from superannuation
  - any other income
- wages/salaries
  - regular overtime
  - commissions and bonuses.

The total income reported on a person’s tax statement should conceptually be quite similar to that reported in the census. However, several reasons are likely to create differences between the two measures. First, and most importantly, many of those who had some income but did not reach the tax-free threshold would not have completed a tax form. Second, the MADIP BLE2011 does not report nontaxable income from sources other than the social security system, such as some scholarships, bursaries, grants and awards. Third, census respondents are asked about their ‘usual’ income in weekly ranges, whereas those who submit a tax form report for the whole year. Fourth, on the tax form, the items listed above are reported individually, and the calculation is made by the Australian Taxation Office. On the census, however, individuals are instructed to do all the aggregation, deduction (or lack thereof) and calculation themselves. This introduces a greater scope for error.

Finally, there are incentives in the tax system to report a relatively low income and strong incentives to report accurately (due to the risk of auditing). There is no incentive to report a high or low income in the census, but there is also no incentive to report accurately. This is particularly the case for individuals on the census form who have their income and other characteristics filled out by someone else in the household.

As noted by the ABS<sup>4</sup> in its description of the BLE2011:

*... annual income reported in the PIT 2010–11 dataset may be inconsistent with annual income reported for the same record in the Census 2011 dataset. This is due to the differing nature in which the PIT 2010–11 information is collected (via a tax return with specific questions about income for tax purposes) compared to the nature in which the Census 2011 information was collected (via a Census form with specific questions about income for statistical purposes).*

To explore the differences and similarities between the income data from the PIT and the income data from the census, we begin with Figure 2, which gives two sets of numbers for the BLE2011 sample by the individual's income category.

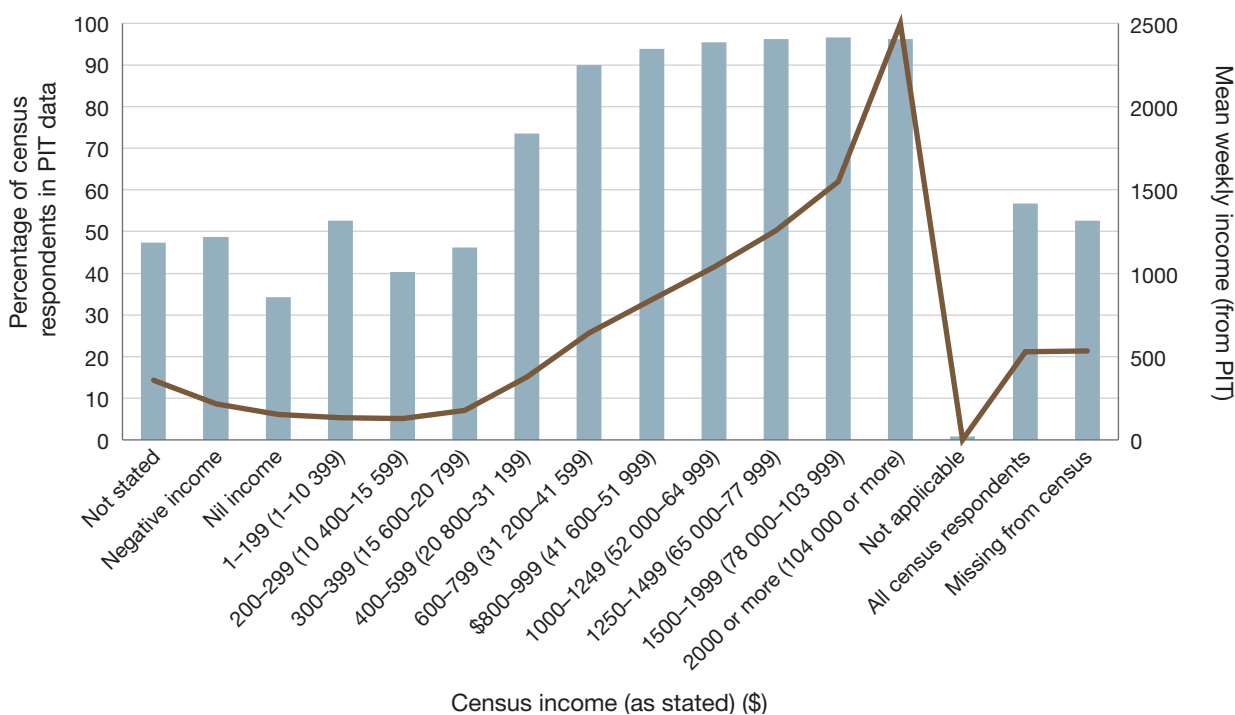
The bars in Figure 2 represent the percentage of those in each income category who are in the PIT data on the BLE2011 (plotted on the left-hand vertical axis). Keeping in mind that the tax-free threshold for the 2010–11 financial year was \$6000, all those in the second income category

(\$200–299 per week) and a large proportion of those in the first category (\$1–199 per week) should be in the PIT data. This was clearly not the case. Only around half or less of those in the first four income categories were in the PIT data, and only a little over 70% of those in the fifth income category (\$400–599 per week) were in the PIT data. Although the percentage increases to around 90% or above for the remaining income categories, there is still a very large number of people not in the PIT data who, according to their census data, should be. At the other end of the distribution (the negative income and nil income categories), there was a non-negligible proportion in the PIT data, when we might expect that to be closer to zero.

This does not necessarily mean that there are errors in either dataset (of either representation or measurement), although there might well be. Rather, it may be that definitional differences – for example, financial year versus weekly – are driving some of the differences.

The second part of Figure 2 is a line graph (plotted against the right-hand y axis) that gives an estimate of the mean weekly total income

**Figure 2 Relationship between census income (2011) and PIT income (2010–11 financial year) in nominal dollars, adults aged 15 years and over in 2011**



PIT = personal income tax

Source: Customised data from the MADIP Basic Longitudinal Extract 2011

from the PIT data based on the person's census income category. For this figure, those not in the PIT data have been assigned an income value of zero. We will return to the validity of this assumption later.

There is very little difference in the mean income (according to PIT data) for those in the first five income categories – all around or well below \$250 per week (with the exception of those who did not state their income in the census). Mean income then increases substantially across the remaining income categories, as we would expect. One of the more important findings from Figure 2 is that, apart from the highest income category, mean income from the PIT data is either at or below the lower bound of the census category. The upper income category is interesting, however. For this group, mean income is around \$2500 per week, which is well above the lower bound of the income category. Given that a number of people in this group have an income below the lower bound (discussed in the next section), this suggests a skewed distribution within that category.

Two other sets of income categories are interesting. On the far left of the graph, we give the average income for those who did not state their income in the census. This is a very large category (around 7.9% of the full 2011 Census, excluding the 'not applicable' category), showing a very high item nonresponse for this question. In previous analysis (e.g. Markham & Biddle 2018), we have excluded this group from our calculation of means and medians, under the implicit assumption that the income for this group is distributed in the same way as for those who state their income. If anything, we have assumed that this is a conservative estimate and that those in this group are likely to have higher incomes (and did not state their income for this reason) than those who stated their income. These assumptions have been made in the absence of any other information, because the census Post Enumeration Survey does not include any income information, or information that would help us predict income (e.g. education, employment status, hours worked, occupation).

The PIT data for those who did not state their income suggests that our assumptions appear to have been incorrect. The mean income for this group (from PIT data) is around \$360 – only a little over half of the mean income for those who did state their income. However, this average should be interpreted with caution, because it is derived from a little under half of the census not-stated sample who are in the PIT data, and from individuals who have been linked either to the Medicare or SSRI data (i.e. excluding individuals with only dummy data whose records have been imputed into the census).

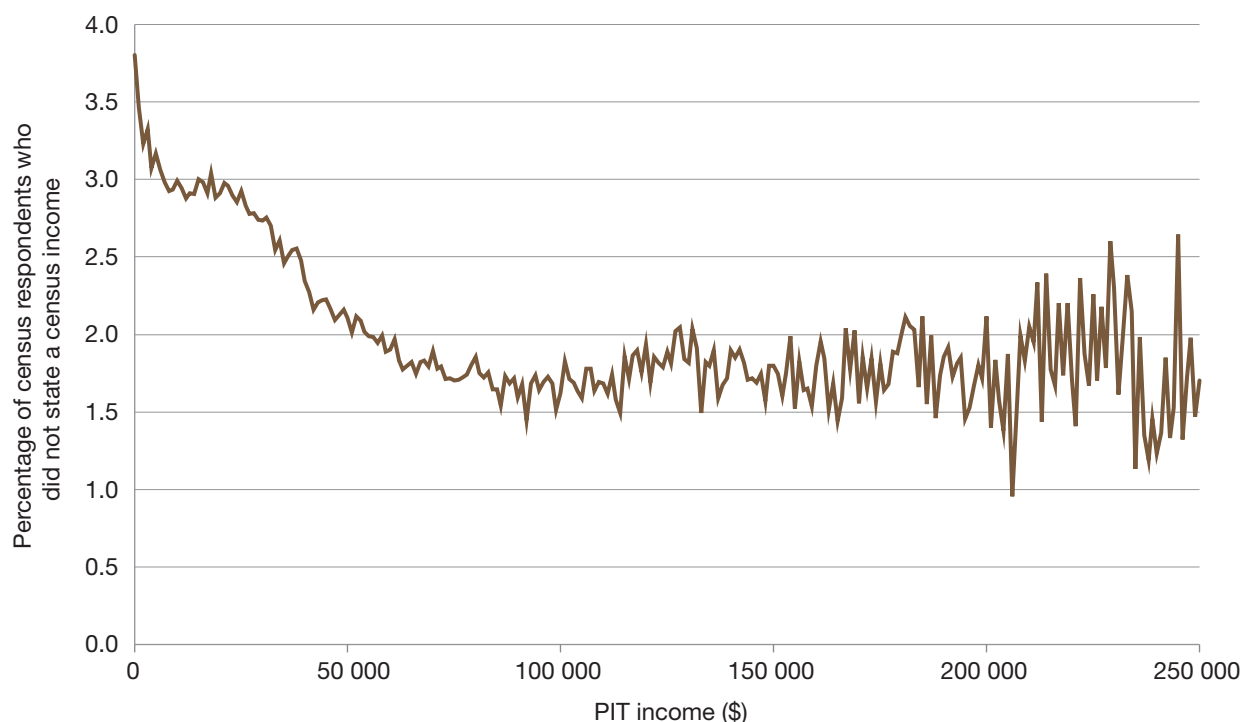
We explore the relationship in more detail in Figure 3, which shows the percentage of people who did not state their income in the census plotted against their income from PIT data. Individuals with zero or negative income in the PIT data are most likely to have not stated their income in the census. The percentage of those who did not state their income in the census declines as PIT-reported income increases across the income distribution until plateauing at around \$70 000–75 000, after which it stays reasonably constant, albeit with some fluctuations due to relatively small samples.

Contrary to our speculation that the not-stated category consists primarily of those with high incomes who do not want to share that information with government, it would appear that the not-stated income category is disproportionately composed of those with relatively low income, or no income at all.

The final two categories in Figure 2 give the PIT data for those who had a linked census record (all the income categories combined) and those who did not have a linked record. The data show that the latter group of individuals were a little less likely to be in the PIT data than the former (56.7% compared with 52.6%), but that the mean income for the two groups was quite similar (within \$5 per week).

We return to the comparison of those who have and do not have a census record in the last subsection of this section. Before then, we consider the factors associated with there being a significant discrepancy between incomes in the census and PIT data.

**Figure 3** Percentage of people with not-stated income in the census across the PIT income distribution (2010–11), adults aged 15 years and over in 2011



PIT = personal income tax

Source: Customised data from the MADIP Basic Longitudinal Extract 2011

### 5.3 What are the predictors of variation between census and PIT income data?

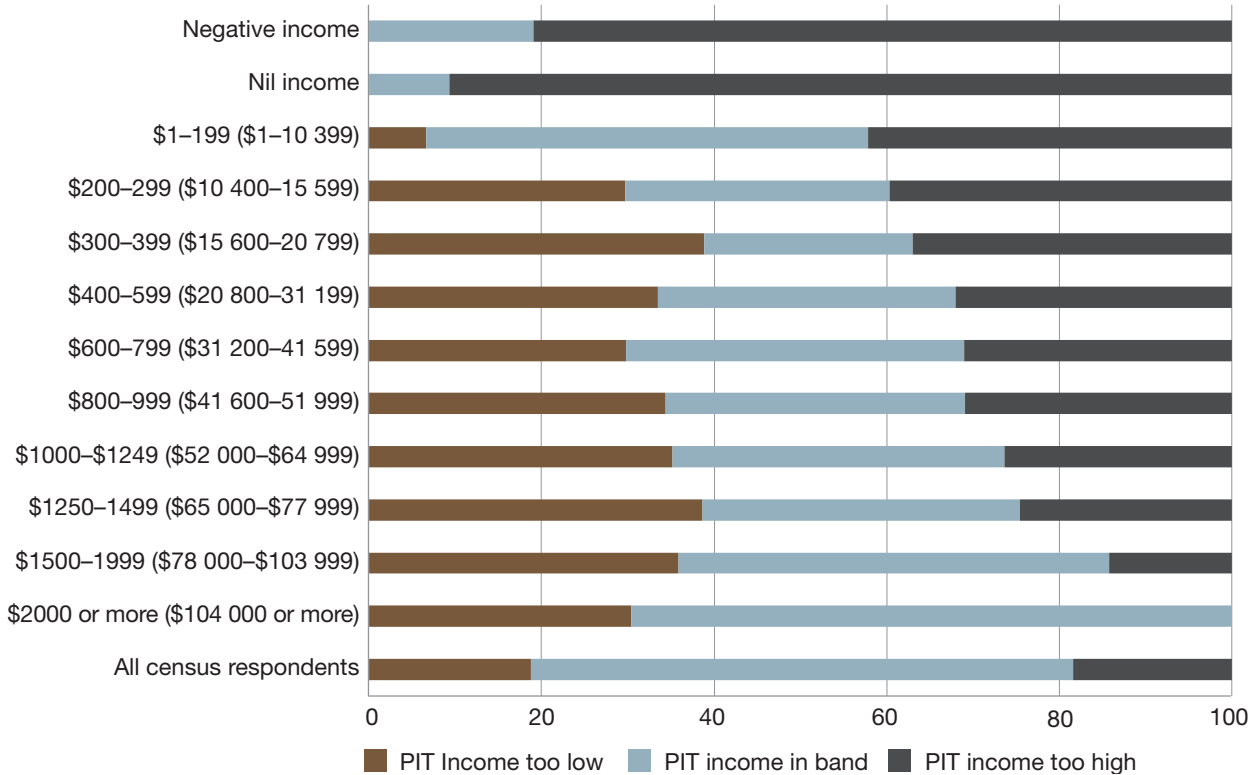
Some valid reasons might explain why the income that a person reports in the census would be different from the income that they report on their tax return. The census asks for a person’s usual income, whereas the tax system incorporates income from across a financial year. Unexpectedly high or low weeks would be included in the summation in the PIT data, but people are unlikely to include them in the census data. The timing is also slightly different, with the PIT data being for the financial year leading up to the census (which ends in June) and the census referring to a particular week (usually the second week in August). However, these differences are minor, and unlikely to be large on average. As shown in Figure 4, however, there is a discrepancy in income for a very large proportion of those for whom we have both PIT and census income data.

To create the data in Figure 4, we first excluded those who did not have a PIT record. By definition, these individuals could not be in the second or third categories as described below. We then broke the remainder of the sample into three categories. The first category, in brown, is those whose income from the PIT data is below the census income category they report. The middle category, in blue, is those whose income from the PIT data falls into their census income category. The final category, in black, is those whose income from the PIT data is above their census category.

Looking at the last bar in Figure 4, the PIT income for most census respondents is in the expected census income category. However, around one-third of the census respondents are not in the expected category; a roughly equal proportion report PIT income above or below their census category.

Not surprisingly, those at the very bottom of the census income distribution either cannot have, or are unlikely to have, too low a PIT

**Figure 4 Percentage of sample whose PIT income is below, within or above their census income category**



PIT = personal income tax  
 Source: Customised data from the MADIP Basic Longitudinal Extract (BLE) 2011

income, whereas those at the top of the income distribution are unlikely or unable to have too high an income. However, in the middle part of the census income distribution – \$200–1500 per week – it is more likely that a person will be outside their census income category rather than in it.

Observable characteristics in the census predict whether or not a person has income from tax data

that is above or below their census income band. We estimated this using a multinomial logit model, with being in the ‘correct’ income category as the base outcome and having an income that is above or below as the alternative categories. We control for age, sex, Indigenous status, migrant status, employment and education. Results are given in Table 4.

**Table 4** Factors associated with whether PIT income is below, within or above census income category

Explanatory variable	PIT lower		PIT higher		PIT in range
	Coefficient	Marginal effect	Coefficient	Marginal effect	Marginal effect
Aged 15–24	–0.087*	–0.034	0.147*	0.045	–0.011
Aged 25–34	0.164*	0.025	0.089*	0.002	–0.026
Aged 35–44 (omitted category)					
Aged 45–54	–0.127*	–0.036	0.093*	0.036	0.000
Aged 55–64	0.116*	–0.002	0.221*	0.039	–0.037
Aged 65–74	0.544*	0.118	0.056*	–0.055	–0.063
Aged 75 plus	0.187*	0.058	–0.146*	–0.054	–0.004
Female	0.126*	0.012	0.128*	0.015	–0.027
Indigenous	0.231*	0.042	0.069*	–0.012	–0.031
Employed	–0.458*	–0.027	–0.737*	–0.116	0.143
Completed year 9 or less	0.028*	–0.014	0.175*	0.038	–0.024
Completed year 10 or 11	–0.006*	–0.009	0.067*	0.016	–0.008
Has a degree	0.009*	0.049	–0.449*	–0.099	0.049
Born in Australia (omitted category)					
Born overseas – arrived 2001–11	0.181*	0.028	0.094*	0.001	–0.029
Born overseas – arrived 1991–2000	0.127*	0.014	0.115*	0.012	–0.026
Born overseas – arrived 1981–90	0.100*	0.007	0.125*	0.018	–0.024
Born overseas – arrived 1980 or earlier	0.013*	–0.006	0.075*	0.016	–0.010
Probability of base case	–0.073	0.302	0.139	0.373	0.325
Sample size	7,331,060				
Pseudo <i>R</i> -squared	0.0169				

PIT = personal income tax

Notes:

1. The base-case person is aged 35–54, male, non-Indigenous and not employed; has completed year 12; does not have a degree; and was born in Australia.
2. Coefficients that are significantly different from 0 at the 1% level of significance are marked with \*.

The most important results in Table 4 are the marginal effects. These are reasonably complicated to interpret, so it is worth stepping through slowly. The first thing to note is that the last line of the table gives the probability of being in a particular category for someone with the base-case characteristics (as described in table note 1). The marginal effect is the difference in probability of being in that category (PIT too low, too high or within range) when we change only one variable and keep all other variables constant. The probabilities (in the last line) add up to 1, whereas the marginal effects add up to 0 (with some rounding).

We can interpret the marginal effects in the last column as the association between that explanatory variable and the probability of having no discrepancy between PIT income and census income. Both the relatively young and the relatively old are less likely to be in the correct category, as are females, the Indigenous population and those with low levels of high-school education. Those who are employed or have a degree are more likely to be in the correct income category, whereas those who were born overseas – particularly recent arrivals – have a lower probability. Indeed, the predictor with the largest marginal effect is employment, a result that may reflect the relatively good alignment of the income concept in the census and PIT for those who are in full-time employment.

Looking at the other two columns of marginal effects, there is a complicated relationship between age and the census/PIT discrepancy. The very young (aged 15–24) are more likely to have a PIT income that is above their census category, but less likely to have one that is below. Those entering their full-time working life (aged 25–34) are more likely to have too low a PIT income compared with their census category, but are no more or less likely to have an income above the threshold. Those towards the end of their prime working years (aged 45–64) are less likely to have a PIT income that is below their census cut-off, but more likely to have a PIT income that is above. The biggest differences, however, are at the very upper end of the age distribution. Those aged 65 years and over (particularly those aged 65–74) are far more likely to have a PIT income that is low relative to the

census, and far less likely to have a PIT income that is high. The prime reason for this is likely to be the receipt of tax-free superannuation income that does not appear in the PIT data but is captured in the census.

A few other characteristics have a directional association with census and PIT income discrepancies. Indigenous Australians are more likely to have a PIT income that is below and less likely to have a PIT income that is above their census category. Those who have not completed year 12 are more likely to have a PIT income that is relatively high and less likely to have a PIT income that is relatively low. Finally, those with a degree are more likely to have a PIT income that is lower than their census category and far less likely to have a PIT income that is higher.

In summary, a number of individual characteristics predict whether there is a discrepancy between PIT data and census data in terms of personal total income. In future analysis, we will add additional information from the SSRI data, as well as a combination of income information from PIT data and census data to obtain a synthetic estimate of total income for individuals and households. Researchers using income measures from one of the data sources only (either the MADIP file or the separate datasets) should keep the above discrepancies in mind when interpreting findings.

## 5.4 What are the characteristics of those with missing census data in the BLE2011?

As mentioned earlier, linkage to the census was far from complete – only 61% of the records in the BLE2011 had a census record. If people from the census who are able to be linked to the Medicare database or the SSRI differ systematically from those who are not, the analysis of census data in the BLE2011 may be biased.

The last two categories in Figure 2 show that there were some differences in whether or not a person was in the PIT database, depending on whether or not they have a linked census record. Those missing from the census had a lower probability of being in PIT data than those who were able to be linked. However, more

promisingly, there were no substantial differences in mean income across samples.

In Table 5, we consider whether there are other differences between those whose census record is available through the BLE2011 and those whose census record is not. Table 5 has three columns of data, based on the source of information. The first column is for those whose census record is in the BLE2011, with the data taken from the BLE2011. This is the sample that is likely to be the focus of most of the analysis of the BLE2011 in the future. The second column gives the results for those in the BLE2011 who do not have a census record. By definition, there is no census-related information for these individuals, so we are restricted to summarising data from other sources in the BLE2011 (age, sex and Indigenous status). The third column comes from outside the BLE2011 and is for the total population from the 2011 Census. Given our focus on using the BLE2011 to study income

and its determinants, we focus on income data or characteristics that are highly predictive of a person's income.

Table 5 shows that there are some differences between the BLE2011 census sample and the rest of the Australian population, but that these differences are not particularly large. Compared with the full census, both groups of the BLE2011 records are younger. This may be a result of grouping and top-coding of age data. However, the gap between the census-linked and non-census-linked records in the BLE2011 is worth noting. In addition, BLE2011 records linked to the census have a slightly lower average income than those not linked to the census, and are more likely to be female, non-Indigenous, not employed, with low levels of education and born in Australia. Indeed, on the basis of the limited available data, it appears that BLE2011 records linked to the census are closer to the full census than BLE2011 records that are not linked to the census.

**Table 5 Demographic and socioeconomic information by census linkage status and for the total census population**

Characteristic	MADIP		
	MADIP census sample	non-census sample	Full census
Average age	37.0	33.9	37.8
Percentage female	0.523	0.471	0.510
Percentage Indigenous	0.024	0.070	0.027
Percentage employed	0.590	na	0.614
Percentage completed year 9 or less	0.151	na	0.142
Percentage completed year 10 or 11	0.351	na	0.338
Percentage completed year 12	0.498	na	0.520
Percentage with a degree	0.199	na	0.211
Percentage born overseas	0.250	na	0.260
Average census income based on midpoints	\$38 590	na	\$39 873

MADIP = Multi-Agency Data Integration Project; na = not applicable

Source: Customised data from the MADIP Basic Longitudinal Extract 2011 (columns 1 and 2); 2011 Census of Population and Housing (column 3)



## 6. Concluding comments and future analysis using MADIP

Understanding the determinants, dynamics and distribution of income within a country is one of the key areas of enquiry for social scientists who have an interest in the economic wellbeing of the residents of that country, either nationally, at a small geographic level or for other population subgroups.

Research to date has sketched a detailed outline of the determinants, dynamics and distribution of income in Australia. We know, for example, that inequality has been somewhat stable during the past few decades, even if the precise distribution of income is still measured with error for a given year. We also know that the inequality of the distribution of income in Australia is somewhere in the middle relative to other OECD countries – lower than for the United States, similar to the United Kingdom and New Zealand, but higher than for most continental European countries. We also know that income is not distributed evenly across Australia: certain regions (mainly outside the capital cities) and certain population subgroups (particularly Indigenous Australians, but also certain migrant cohorts) have lower average incomes and lower incomes than would be predicted based on their human capital characteristics.

However, we know relatively little about a number of aspects of income in Australia. These include detailed information on the spatial dimensions of income inequality, income mobility in and out of the top of the distribution, sociodemographic characteristics at the top of the distribution, the top of the household income distribution, the impact of local geographic inequality and other characteristics on income change, and the complete income distribution of small subpopulations of particular policy interest. The opening up of access to data from the MADIP

BLE2011 provides an opportunity to create new knowledge in these areas.

The aim of this paper was to begin the external validation of these data, with a particular focus on what the data can tell us about the distribution, dynamics and determinants of income in Australia. We found that:

- we have tax data on slightly over half of the BLE2011 population
- if a person has tax data for 1 year, they are likely to have it for multiple years
- income from PIT data correlates highly with income from the census (coefficient of 0.7791) but around one-third of the sample has an income from PIT data that is either above or below their census income category (in a roughly even split)
- there are consistent predictors for being outside that range
- people with linked census information in the BLE2011 have similar, but not completely identical, characteristics to those in the full 2011 Census; this validation is useful not only for analysis of the BLE2011 but also for researchers who are considering analysing income on the PIT databases or 2011 Census in their unlinked forms.

Based on these results, our view is that the BLE2011 has the potential to shed new light on the determinants, dynamics and distribution of income in Australia. However, analysis of the dataset should be carried out with caution, taking into account some of the limitations outlined above. As with any new source of data, it is important to share insights and learnings from this data source across a community of users, and we encourage productive collaborations and exchange of knowledge across all points of the research cycle.



## Notes

1. <https://alife-research.app>
2. [www.abs.gov.au/ausstats/abs@.nsf/Lookup/1700.0main+features110Australia](http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1700.0main+features110Australia)
3. [www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Standard+Geographical+Classification+\(ASGC\)](http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Standard+Geographical+Classification+(ASGC))
4. [www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1700.0~Australia~Main%20Features~Methodology~120](http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1700.0~Australia~Main%20Features~Methodology~120)

## References

- Abello A & Harding A (2006). Income mobility and financial disadvantage: Australian children. *Agenda* 13(1):31–48.
- Athanasopoulos G & Vahid F (2003). Statistical inference and changes in income inequality in Australia. *Economic Record* 79(247):412–424.
- Atkinson AB & Leigh A (2005). *The distribution of top incomes in Australia*, ANU Centre for Economic Policy Research Discussion Paper 514, Australian National University, Canberra, <https://openresearch-repository.anu.edu.au/handle/1885/43259>.
- & Leigh A (2007). The distribution of top incomes in Australia. *Economic Record* 83(262):247–261.
- Barrett GF, Crossley TF & Worswick C (2000). Consumption and income inequality in Australia. *Economic Record* 76(233):116–138.
- Beeson PE (2017). Agglomeration economies and productivity growth. In: Mills ES & McDonald JF (eds), *Sources of metropolitan growth*, Routledge, 19–33.
- Biddle N & Montaigne M (2017). Income inequality in Australia: decomposing by city and suburb. *Economic Papers* 36(4):367–379.
- , Gray M & Schwab RJ (2017). *Measuring and analysing success for Aboriginal and Torres Strait Islander Australians*, Working Paper 122, Centre for Aboriginal Economic Policy Research, Australian National University, Canberra, [http://caepr.cass.anu.edu.au/sites/default/files/docs/Working\\_Paper\\_122\\_2017.pdf](http://caepr.cass.anu.edu.au/sites/default/files/docs/Working_Paper_122_2017.pdf).
- Buddelmeyer H & Verick S (2008). Understanding the drivers of poverty dynamics in Australian households. *Economic Record* 84(266):310–321.

- Burkhauser RV, Hahn MH & Wilkins R (2015). Measuring top incomes using tax record data: a cautionary tale from Australia. *Journal of Economic Inequality* 13(2):181–205.
- Chipperfield J, Brown JJ & Watson N (2017). The Australian Census Longitudinal Dataset: using record linkage to create a longitudinal sample from a series of cross-sections. *Australian & New Zealand Journal of Statistics* 59(1):1–16.
- Claessens S, Dell’Ariccia G, Igan D & Laeven L (2010). Cross-country experiences and policy implications from the global financial crisis. *Economic Policy* 25(62):267–293.
- Ellis L & Andrews D (2001). *City sizes, housing costs, and wealth*, Research Discussion Paper 2001–08, Reserve Bank of Australia, Canberra, <https://www.rba.gov.au/publications/rdp/2001/pdf/rdp2001-08.pdf>.
- Finlay R (2012). *The distribution of household wealth in Australia: evidence from the 2010 HILDA survey*, Bulletin (March Quarter 2012), Reserve Bank of Australia, Canberra.
- Fleming DA & Measham TG (2015). Income inequality across Australian regions during the mining boom: 2001–11. *Australian Geographer* 46(2):203–216.
- Gaston N & Rajaguru G (2009). The long-run determinants of Australian income inequality. *Economic Record* 85(270):260–275.
- Greenwood J, Guner N, Kocharkov G & Santos C (2014). Marry your like: assortative mating and income inequality. *American Economic Review* 104(5):348–353.
- Harding A (1995). The impact of health, education and housing outlays upon income distribution in Australia in the 1990s. *Australian Economic Review* 28(3):71–86.
- (1997). The suffering middle: trends in income inequality in Australia, 1982 to 1993–94. *Australian Economic Review* 30(4):341–358.
- Headey B, Marks G & Wooden M (2005). The dynamics of income poverty in Australia: evidence from the first three waves of the HILDA survey. *Australian Journal of Social Issues* 40(4):541–552.
- Hunter B (2003). Trends in neighbourhood inequality of Australian, Canadian, and United States of America cities since the 1970s. *Australian Economic History Review* 43(1):22–44.
- & Gregory RG (1996). An exploration of the relationship between changing inequality of individual, household and regional inequality in Australian cities. *Urban Policy and Research* 14(3):171–182.
- Johnson D & Wilkins R (2004). Effects of changes in family composition and employment patterns on the distribution of income in Australia: 1981–1982 to 1997–1998. *Economic Record* 80(249):219–238.
- Kaplan G, Cava GL & Stone T (2018). Household economic inequality in Australia. *Economic Record* 94(305):117–134.
- Leigh A (2005). Deriving long-run inequality series from tax data. *Economic Record* 81(s1):S58–S70.
- (2006). Trust, inequality and ethnic heterogeneity. *Economic Record* 82(258):268–280.
- Markham F & Biddle N (2018). Income, poverty and inequality, 2016 Census Paper 2, Centre for Aboriginal Economic Policy Research, Australian National University, Canberra, <http://caep.cass.anu.edu.au/research/publications/income-poverty-and-inequality>.
- OECD (Organisation for Economic Co-operation and Development) (2010). *OECD economic surveys: Australia 2010*, OECD Publishing, Paris.
- Productivity Commission (2018). *Rising inequality? A stocktake of the evidence*, Research Paper, Productivity Commission, Canberra, <https://www.pc.gov.au/research/completed/rising-inequality>.
- Rohde N, Tang KK & Rao P (2010). *Income inequality, mobility and economic insecurity in Australia*, Discussion Paper 407, School of Economics, University of Queensland, Brisbane.
- Saunders P (1993). Longer run changes in the distribution of income in Australia. *Economic Record* 69(4):353–366.
- (2004). Examining recent changes in income distribution in Australia. *Economic and Labour Relations Review* 15(1):51–73.

- & Bradbury B (2006). Monitoring trends in poverty and income distribution: data, methodology and measurement. *Economic Record* 82(258):341–364.
- Solon G (1989). Biases in the estimation of intergenerational earnings correlations. *Review of Economics and Statistics* 71(1):172–174.
- Summerfield M, Bevitt A, Freidin S, Hahn M, La N, Macalalad N, O’Shea M, Watson N, Wilkins R & Wooden M (2017). *HILDA user manual: release 16*, Melbourne Institute of Applied Economic and Social Research, Melbourne.
- Venn D & Hunter B (2018). *Poverty transitions in non-remote Indigenous households: the role of labour market and household dynamics*, Working Paper 124, Centre for Aboriginal Economic Policy Research, Australian National University, Canberra.
- Whiteford P (2010). The Australian tax-transfer system: architecture and outcomes. *Economic Record* 86(275):528–544.
- Wiesel I (2018). *Power, glamour and angst: inside Australia’s elite neighbourhoods*, Palgrave Macmillan, Singapore.
- Wilkins R (2014a). Evaluating the evidence on income inequality in Australia in the 2000s. *Economic Record* 90(288):63–89.
- (2014b). The distribution and dynamics of household income. In: Wilkins R (ed.), *Families, incomes and jobs*, vol. 9, *A statistical report on waves 1 to 11 of the Household, Income and Labour Dynamics in Australia survey*, Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Melbourne, 22–31.
- (2015). Measuring income inequality in Australia. *Australian Economic Review* 48(1):93–102.
- (2016). *The Household, Income and Labour Dynamics in Australia survey: selected findings from waves 1 to 14*, Melbourne Institute of Applied Economic and Social Research, Melbourne, 124.
- World Bank (2018). *World Development Indicators*, World Bank, Washington, DC.



## CENTRE FOR SOCIAL RESEARCH & METHODS

+61 2 6125 1279

[csrm.comms@anu.edu.au](mailto:csrm.comms@anu.edu.au)

The Australian National University  
Canberra ACT 2601 Australia

[www.anu.edu.au](http://www.anu.edu.au)

CRICOS PROVIDER NO. 00120C