Australian Government
Productivity Commission

BLADE for productivity research

Productivity Commission
Staff Working Paper

May 2021

Henry McMillan
Colin Burns

The views expressed in this paper are those of the staff involved and do not necessarily reflect the views of the Productivity Commission.

ISBN 978-1-74037-717-1 (PDF)

An appropriate reference for this publication is:

McMillan, H. and Burns, C. 2021, *BLADE for productivity research*, Productivity Commission Staff Working Paper, May.

Publications enquiries

Media, Publications and Web, phone: (03) 9653 2244 or email: mpw@pc.gov.au

The Productivity Commission

The Productivity Commission is the Australian Government's independent research and advisory body on a range of economic, social and environmental issues affecting the welfare of Australians. Its role, expressed most simply, is to help governments make better policies, in the long term interest of the Australian community.

The Commission's independence is underpinned by an Act of Parliament. Its processes and outputs are open to public scrutiny and are driven by concern for the wellbeing of the community as a whole.

Further information on the Productivity Commission can be obtained from the Commission's website (www.pc.gov.au).

Contents

Preface

The results of this study is based, in part, on ABR data supplied by the Registrar to the ABS under *A New Tax System (Australian Business Number) Act 1999* and tax data supplied by the ATO to the ABS under the *Taxation Administration Act 1953*. These require that such data is only used for the purpose of carrying out functions of the ABS. No individual information collected under the *Census and Statistics Act 1905* is provided back to the Registrar or ATO for administrative or regulatory purposes. Any discussion of data limitations or weaknesses is in the context of using the data for statistical purposes, and is not related to the ability of the data to support the ABR or ATO's core operational requirements. Legislative requirements to ensure privacy and secrecy of this data have been followed. Only people authorised under the *Australian Bureau of Statistics Act 1975* have been allowed to view data about any particular firm in conducting these analyses. 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

BLADE for productivity research

Key points

- The Business Longitudinal Analysis Data Environment (BLADE) is a financial census of almost all Australian businesses spanning 2001-02 to 2018-19. It contains key variables on turnover, employment, wages, debt and depreciation.
- BLADE is a unique database in Australia, providing information on firm activity not available through any other channels. It is Australia's best resource for firm-level productivity research and provides more depth to Australia's macro productivity narrative.
- BLADE-based productivity estimates have some conceptual limitations. These mean that macro productivity estimates must impute or proxy other variables. Specifically:
  - BLADE contains no separate price and quantity measures
  - BLADE contains no direct estimates of the firm's capital stock
  - assumptions required for macro productivity calculation (e.g. positive value add) do not always hold at the firm level.
- Together this means that point estimates of productivity at the firm level are still in their infancy. Point estimates of firm-level labour productivity are currently more difficult to compare over time due to lack of firm price deflators.
- Existing research attempts to mitigate these limitations with various assumptions and modelling approaches.
- BLADE is a valuable resource for competition analysis, entry and survival, disaggregation of macro business trends, as well as understanding industry characteristics not easily available in other datasets. Many of BLADE's other applications can be interpreted at face value without extensive modelling.
- A key challenge for all administrative datasets lies in the economic interpretation of data that is collected for administrative purposes. Seemingly simple concepts such as 'revenue' or 'profit' have different meanings in the contexts of accounting and economic analysis.
- BLADE poses technical hurdles for prospective researchers — the volume, administrative format and many non-reported values can bog down even simple analysis.
  - Many of the variables in BLADE that would be of interest to researchers (for example those taken from surveys) are not reported for a large portion of firms and either have to be imputed or else lose a large portion of the observations.
  - Components of the dataset (like many linked datasets) need to be merged by individual researchers. In practice, this can lead to researchers using a fraction of the available data.
- It is always important for researchers to be clear about their assumptions and understand their data. As BLADE is applied to more questions, researchers will build a greater understanding of its strengths and limitations.

This paper provides an overview of the Business Longitudinal Analysis Data Environment (BLADE), its potential uses as a research tool and its limitations. The overview is based on examination by Commission staff during 2019 of the potential use of BLADE to inform Commission analysis.

BLADE is a valuable resource for economic research. BLADE’s strength lies in being, in effect, a census of financial information on all Australian firms between 2001-02 and 2018-19 (ongoing) as reported to the Australian Tax Office (ATO), with much of this data also linked to important ABS economic surveys. This dataset has several advantages over alternatives.

- Compared with ABS industry productivity data, firm level data provide more precise analysis on how individual businesses are affected differently by policies or economic conditions. For example, some businesses may participate in a government program and others may not, and some businesses may be affected by prudential constraints on bank lending more than others. It also allows an understanding of the effects of firm entry and exit.
- Compared with ABS business surveys, BLADE has greater coverage and scope. And many key ABS businesses surveys are themselves linked to BLADE.
- Compared with commercial datasets drawn from public company annual reports, BLADE has much broader coverage including small private companies and unincorporated businesses. The linkage to ABS surveys also provides data on drivers of productivity growth (such as innovation, R&D, management capability) not available from annual reports.
- Compared with custom surveys, ABS surveys have much higher response rates as businesses are legally compelled to respond.

On the other hand, BLADE also has some limitations. For example, with 135 million observations split over 71 separate files, it can be unwieldy, and researchers generally only use a fraction of the available data. Some important variables do not exist for either a large portion of firms (for example, employment is missing for many firms, some of whom may employ people in reality; survey variables are only present for those firms included in the survey). Being mostly tax data, with no quantity or price measures, the dataset has limitations for productivity analysis or calculating markups. These issues are discussed in more detail below.

Overview

BLADE has been funded in recent years by the Data Integration Partnership for Australia, while early development was funded by the Department of Industry, Science, Energy and Resources (DISER). Development priorities are proposed by the ABS and discussed with the BLADE Technical Advisory Group.

BLADE consists of two parts: a ‘core’ component made up of ATO data, and various ‘modules’ that are a mix of ABS business surveys and administrative data from other agencies (such as IP Australia).

The Core

BLADE core contains tax information from:

- Business Activity Statements (BAS, a form used to collect GST, which focusses on current revenues and expenses)
- Business Income Tax statements (BIT, a form used to collect business income tax, which includes additional balance sheet information)
- Pay As You Go statements (PAYG, used to withhold income tax from employees).

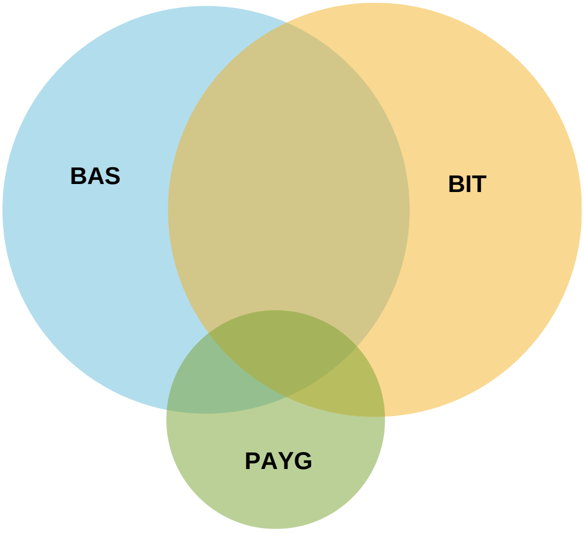
These tax data are merged through the ABS business register (FRAME) that provides demographic data such as geography, firm creation date and business group identifiers.

The coverage of each of these tax datasets varies (figure 1). By far the most complete dataset is the BAS, which covers most active firms, followed by BIT and then PAYG. There has also been a tendency for the coverage of the BAS to increase over time (possibly reflecting a fixed nominal reporting threshold since 2008), with the other two tax dataset’s proportional coverage remaining about the same (table 1). It is worth noting that the FRAME column in table 1 does not represent the number of firms that would be expected in any given year, but can be thought of in a loose sense as a representation of the total number of firms across all years (from the perspective of that given year).

For the reasons discussed below, most research to date has only used a subset of BLADE core, typically including only one or two of the three datasets, or restricting their analysis to only a few years’ worth of data.

Figure 1 **Some variables are recorded only for a minority of firms**

Proportion of firms contained in BAS, BIT and PAYG[a] (and overlaps) 2017-18



[a] BAS = Business Activity Statement, BIT = Business Income Tax statement, PAYG = Pay As You Go statements.

Source: Commission estimates using the Business Longitudinal Analysis Data Environment (BLADE).

Table 1 **Number of firms by dataset**

Financial year ending	*BAS*	*BIT*	*PAYG*	*FRAME*
2002	2 327 859	2 112 496	776 132	9 313 206
2003	2 368 882	2 156 642	796 641	9 313 206
2004	2 423 599	2 205 959	815 504	9 306 423
2005	2 488 000	2 262 039	700 924	9 304 979
2006	2 527 045	2 274 603	735 337	9 302 270
2007	2 615 173	2 198 745	714 648	9 299 017
2008	2 641 391	2 308 723	739 814	9 297 507
2009	2 610 531	2 192 816	740 820	9 294 881
2010	2 626 753	2 371 421	739 703	9 292 534
2011	2 647 391	2 327 262	742 966	9 290 727
2012	2 635 631	2 354 279	743 163	9 288 779
2013	2 610 828	2 395 962	746 587	9 287 540
2014	2 611 070	2 466 064	764 795	9 285 988
2015	2 609 961	2 480 758	786 929	9 284 562
2016	2 664 696	2 734 614	771 629	9 283 612
2017	2 707 977	2 798 271	798 445	9 281 710
2018	2 726 096	2 815 139	786 085	9 280 209
2019	2 570 422	-	731 132	9 279 772

Source: Commission estimates using the Business Longitudinal Analysis Data Environment (BLADE).

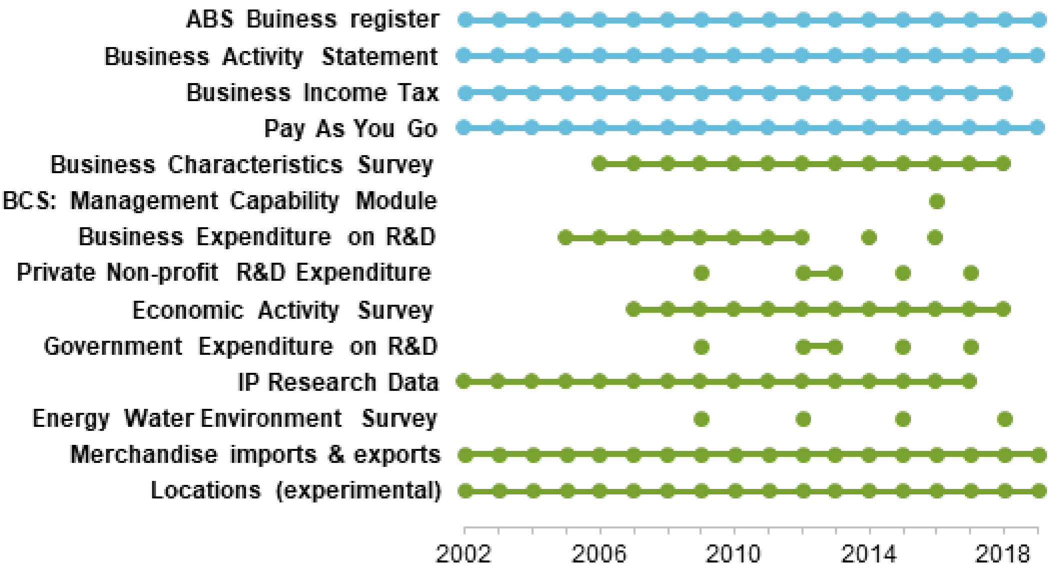
Modules

BLADE also contains 'modules' which link the core data to ABS surveys or administrative datasets from other agencies, including:

- Business Characteristics Survey (BCS)
- Business Characteristics Survey: Management Capability Module (BCSM, 2016 only)
- Business Expenditure on Research & Development (BERD)
- Private Non-profit Expenditure on Research and Development (NERD)
- Economic Activity Survey (EAS)
- Government Expenditure on Research & Development (GERD)
- Intellectual Property Longitudinal Research Data (administrative data from IP Australia)
- Merchandise imports and exports
- Energy, Water and Environment Survey
- Locations (SA1 level; experimental) (figure 2).

Many of BLADE's modules are not yet widely used in literature — particularly as more are added with each BLADE revision.

Figure 2 **BLADE module coverage**



Source: ABS BLADE data item list (unpublished.

BLADE's capabilities

Types of data

BLADE core contains longitudinal administrative tax data on every GST-paying firm between 2001-02 and 2018-19 (ongoing). The main variable types include:

- income statement information, such as revenues (split between income from domestic sales and exports), expenses (for example wages, depreciation and interest)
- balance sheet information, such as assets, liabilities etc.
- employment for some firms, including the employee headcount and the number of full-time equivalents (FTE)
- miscellaneous information such as foreign ownership, use of R&D tax incentives, and location of the main office.

Appendix B provides more detailed information on the specific variables contained within each dataset.

The BLADE modules contain much more information. The ABS surveys (such as BCS, and EAS) follow only a rotating subset of firms, while the other administrative datasets (such as the IP Australia data) have much broader coverage.

Researchers may use modules if the sample is adequate for their projects. They may also be useful for value imputation. That is, when imputing values not available in BLADE core, these more granular surveys could be used to verify output (or as train/validate/test datasets for imputation).

Potential uses

BLADE has numerous potential research uses:

- *Productivity analysis*: despite the dataset lacking information on prices, volumes and capital stocks, many papers have attempted to estimate firm-level productivity. This has included: labour productivity, multifactor productivity and analysis of the dispersion of productivity over time.
- *Program evaluation*: users can link administrative data on which businesses participate in government business support programs with BLADE datasets. This can allow a before and after assessment of growth in sales, employment and productivity.
- *Characteristics of small businesses and sole traders*: BLADE contains far more information on sole traders and small businesses than most datasets of its kind. This allows much more detailed analysis of the (longitudinal) characteristics of small firms and sole traders.

- *Profitability and firm growth*: many papers have looked at the growth of firm turnover and profitability over time.
- *Firm entry and exit*: BLADE contains granular information on firm entry and exit. For example, some research has looked at the changing characteristics of firms entering and exiting, the effect this has on productivity and the effect this has on competition.
- *Competition and concentration*: because turnover is known for almost all firms in a particular industry, concentration can be measured at a granular level (four digit industry definition and SA1 geography). BCS also contains subjective judgements of managers on the degree of competition in a market. Some research has also attempted to estimate the markups at a firm level over time.
- *Export and foreign ownership*: BLADE contains information on foreign ownership, debt, export sales and detailed information on specific merchandise imports and exports.
- *Wages and labour share of income analysis*: because information on the total wage bill and the number of full-time equivalent employees is known for many firms both average wages and the labour share of income can be estimated at a firm level.
- *Availability of credit*: the BCS contains information on whether firms applied for and were able to obtain credit. This can be linked to data on financial health (debt to equity and other ratios), management characteristics etc.
- *Financial health*: BLADE has the required variables to monitor key debt and income ratios which can help identify firms unable to meet their debt obligations.
- *Uptake and effect of intellectual property*: given that most corporate IP information is contained in BLADE, which can be linked with important financial metrics.
- *Innovation activities other than IP*: BCS and BERD asks firms numerous questions about innovation and R&D activity.
- *Custom data integration*: researchers can apply to create a data integration project which links BLADE to some other useful dataset. For example, ABARES has linked BLADE to farm data to gauge financial impacts of changes in weather (ABARES 2019).

Appendix A provides an overview of BLADE research as a guide to which organisations are using which approaches.

Better coverage of sole traders than most firm-level datasets

BLADE covers all GST paying firms. As all firms with GST turnover (gross income minus GST) of $75 000 per year or more must report GST (ATO 2021), the BLADE dataset contains a greater number of small firms (especially sole traders) than comparable overseas datasets. For example, the United States' *Longitudinal Business Database* only covers businesses with paid employees (United States Census Bureau 2021).

Conceptual limitations

Few volume and no price measures

Almost all core BLADE information consists of accounting values, collected for tax purposes. This means that there is little information on the physical quantities or prices of outputs and inputs. The sole exception to this is the information on employee headcount[1] in the PAYG summary, which can be used to calculate average wages per employee (or other variables on a per employee basis).

This absence of price or volume information makes BLADE difficult to use for certain purposes, such as estimating firm-level productivity or mark-up. Usually, in order to estimate productivity changes over time, prices of output must be held constant to prevent non-productivity related issues from affecting measures of output. For example, if one does not hold prices constant then changes in output (measured in terms of revenue minus intermediate input costs) may reflect changes in consumer demand or the degree of competition in the market.

There is a literature which examines the implications of using revenue as an output measure for productivity estimation, and shows output- and revenue-based measures of productivity are correlated (Foster, Haltiwanger and Syverson 2008). However, the conceptual issue remains, since revenue is *by definition* correlated with output (since revenue equals price times output). Moreover, in some papers, physical productivity has found to be inversely related to output price (Foster, Haltiwanger and Syverson 2008).

Some researchers use aggregate price indices (for example, at the division level) to deflate firm level revenues (Andrews and Hansell 2019) but this will only be the equivalent of deflating by firm-level prices if all firms in the industry provide single products that are perfectly substitutable. That is, under perfect competition. If goods are not perfect substitutes in a particular industry, then aggregate price deflators may not reflect important firm-specific factors such as market power, or changes in the mix of goods that the firm provides (industries are normally defined broadly enough to encompass multiple distinct products). Aggregate price indices might be appropriate if differences in firm prices reflect only quality differences, such that, on a quality-adjusted basis, goods in the same industry are perfect substitutes. But this will only be appropriate for single output firms under a narrow industry definition.

Multifactor productivity

Estimating multifactor productivity in BLADE is complicated by several factors: the absence of prices, or measures of physical units (above), the absence of measures of capital and a potential for negative value add at the firm level. However, BLADE does contain detailed information on investment and the tax deductions for depreciation over a reasonable time

[1] BLADE also has a variable, Full-time equivalent employees (FTE) that the ABS imputes using the wages and headcount variables in the PAYG dataset (Hansell, Nguyen and Soriano 2015).

frame, meaning researchers could estimate capital via perpetual inventory methods. As of the end of 2019, the ABS have also added termination values of tangible and intangible assets opening new avenues for capital calculation.[2]

Modern productivity analysis at an industry level normally begins by accumulating investment data over long periods of time and assuming rates of depreciation for each asset class in order to produce estimates of a net capital stock.[3] This approach is called a 'perpetual inventory model' (ABS 2015, p. 361).

While perpetual inventory calculations are theoretically possible in BLADE, the practical difficulties are greater than for industry estimates. In particular, while industry estimates in the national accounts use investment data spanning over 60 years, the BLADE dataset contains only 16 years of data. Firm entries, exits, mergers and data non-reporting mean the effective timeframe for most businesses is even shorter. Moreover, the investment data does not differentiate between different types of capital (as in the National Accounts), so different assets cannot be given different weights (via the rates of return) and different rates of depreciation as is standard practice.

Output measures may also limit the ability to calculate multifactor productivity. While, in aggregate, industries will have positive value added, this does not hold at the firm level. Negative value added can occur where startups are developing their business model, due to adverse shocks such as drought, due to financial transfers between related businesses, due to poor business outcomes or due to data misreporting. Many firms also lack labour and investment data.

Prof. Keven Fox, is currently leading research into a systematic approach to the imputation and output measurement issues surrounding capital and multifactor productivity estimation in BLADE. Prof. Fox's work and code base was set for release in 2019 but has been delayed (Fox 2019).

Identifying a business unit

Ideally, researchers would be able to identify *every* firm in a way that distinguishes between business arms in different industries, that is, an economic firm definition rather than a legal classifier. In practice, this is rarely the case. BLADE suffers from the same difficulties in defining a business unit that plague most other economic statistics, including the national accounts. Many firms are multi-industry conglomerates that need to be subdivided for the purposes of economic statistics. Even relatively simple firms may operate across multiple industries. For example, a computer repair shop may primarily engage in computer repairs (part of 'other services') but also sell computer parts (part of 'retail trade'). Ideally, the

[2] These new variables for termination values of intangible and other depreciating assets are c_terindep, p_terintas, t_terintas, i_terdepas, c_terothde, p_terothas, t_terothas and i_terothas

[3] Assumptions about the user cost of capital are also made to allow aggregation of different asset types into a measure of 'capital services'.

researcher would be able to apportion the output and inputs of the business between its various activities, but usually statistical agencies classify businesses by the activity that constitutes the largest portion of its value add (ABS 2006, p. 21).

In BLADE, firms are measured by Australian Business Number (ABN), which is limited in a few key features:

- not every ABN is present in every dataset
- networks of ABNs can obscure the boundaries and functions of firms
- ABN data does not identify mergers and acquisitions in a clear way.

The ABS solution to some of these problems is a scheme of detailed profiling of larger firms to determine 'type of activity units' or TAUs. TAUs are designed to measure an economic producer, rather than a legal entity or ABN. TAUs are defined by type of production and can have n:n relationships with ABNs. All firms which are not profiled are assumed to have a 1 to 1 relationship with ABN.

Though TAUs aid in firm identification, they suffer from a common BLADE problem: incomplete coverage. Profiling organisations for TAUs is a tedious, manual job and it would be impractical for the ABS to fully profile the Australian economy.

Tax data vs financial data vs management data vs economic data

Large companies calculate profits differently for different purposes. Reported profits for tax purposes may not align with profits for financial reporting purposes, which in turn may not align with books kept for internal managerial purposes. There can also be important differences between financial data and economic concepts. For example, depreciation calculated for financial reporting need not reflect economic depreciation (i.e. the fall in usefulness of an asset over time). In addition, tax rules surrounding depreciation may lead to deductable depreciation being brought forward.

These complications mean that correspondences can be missing between a variable in BLADE and the economic phenomena the researcher is investigating.

How well does microdata fit with aggregate national accounts data?

Despite the potential discrepancies between data used for tax administration, and their interpretation to represent economic concepts, most research indicates that firm-level administrative data do track national accounts data relatively well.

Unpublished work from the Treasury and Australian Bureau of Statistics has compared BLADE aggregates to national accounts supply use tables.[4] BLADE variables — such as

[4] For access to these materials, please contact David Hansell (david.hansell@treasury.gov.au).

total sales and non-capital purchases — were used to proxy output, intermediate use and value added. Most BLADE industry divisions aggregates showed substantial correlation in level comparisons but growth rates were far less correlated. Almost all industry divisions presented an output level correlation higher than 0.9 and value added higher than 0.8. However, the analysis also found that finance and public administration did not track national accounts well.

Kevin Fox's research — to be published, presented at the 2019 Economic Measurement Group conference — compared industry level estimates of multifactor productivity obtained using BLADE to ABS national accounts data and found they were broadly in line (though there were anomalies, such as finance, with weak correlation across the two datasets) (Fox 2019).

Technical limitations

Administrative formats and difficulty handling the size of the data

The dataset as given by the ABS are a series of comma separated value (CSV) files; one for each financial year and for component of the core and modules that must be combined and cleaned before use. This process of combining and cleaning the datasets is time consuming and is so taxing on computation power that most researchers cannot even use the full core datasets of BLADE, and instead use an abridged version of core BLADE with fewer variables (referred to as 'baby BLADE').

In practice, a full outer join of BLADE may not fit within a Datalab (which the ABS uses to host BLADE) server memory (100 to 200GB) and will not fit when using software such as Stata or Eviews. Out of memory solutions are also difficult as Python and R often rely on external database software such as SQL or Apache Spark for large data. SAS provides one out of memory solution, however, Datalab storage suffers from slow disk transfer speeds which can push out execution times.

As of 2020, the ABS has begun transitioning Datalab projects to new cloud-based infrastructure to provide greater flexibility in analytical tools and server scaling for large data projects such as BLADE. As the rollout continues, researchers will be able to provide better solutions to BLADE data management.

An additional issue is that the ABS, admirably, produces regular updates to the BLADE dataset, and the accompanying changes in formats can break existing code. For example, in past releases, FRAME only contained active firms but now it contains every firm from every time period in every financial year dataset.

Missing data

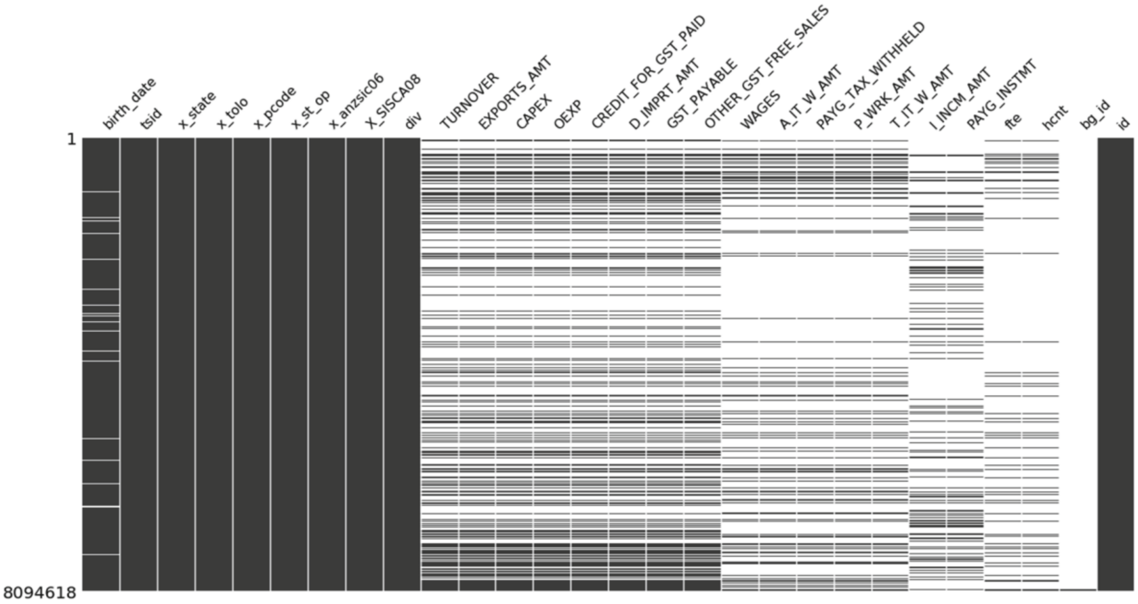
Many important variables are missing for a very large number of observations (figure 3). This does not necessarily mean that the value is genuinely missing — firms are not liable to complete every section of every tax form. Most of the missing data is because the BIT and PAYG datasets do not cover all BAS submitting firms, and so all the variables contained in these (such as employee headcount and firm assets) are missing for many firms. Even where a firm has completed all three tax forms, there are often cases of them omitting a particular variable, or setting it as zero where this does not make sense (eg. reporting their wage bill as zero when they report a positive employee headcount).

In many cases, even though a firm is not obligated to fill out a particular tax form (such as BIT or PAYG), this does not mean that the relevant information does not exist for that firm. For example, even though a firm may not have to fill in a BIT form, the firm will likely still have revenue, profits, assets and liabilities but the researcher will have no information about these. In this sense, even though the tax data does not exist, from the researchers perspective the information is ‘missing’ for that firm.

There is little consensus on how to deal with these issues of missing variables in BLADE or other datasets (Little and Rubin 2019). Some researchers have tended to omit observations with missing variables while others have attempted to impute them (Suresh et al. 2019). Keven Fox (2019) has demonstrated the need for a more comprehensive approach to how missing variables and other common problems (e.g. negative value add) are handled in the construction of BLADE productivity statistics. At time of presentation, Fox intended the release of the framework to harmonise how researchers approach these common problems.

Figure 3 **Gaps in key BLADE variables**

Sparsity denotes the extent of missing entries for each variable



[a] **birth_date** = date of firm creation; **tsid** = time series id (date); **x_state** = state; **x_tolo** = type of legal organisation; **x_pcode** = postcode; **x_st_op** = states of operation; **x_anzsic06** = 4 digit ANZIC code; **X_SISCA08** = SISCA industry sector; **div** = industry division; **TURNOVER** = turnover; **EXPORTS_AMT** = export sales; **CAPEX** = capital expenditure; **OEXP** = non-capital expenditure; **CREDIT_FOR_GST_PAID** = GST on purchases; **D_IMPRT_AMT** = Imported goods with GST deferred; **GST_PAYABLE** = GST on sales; **OTHER_GST_FREE_SALES** = GST-free sales; **WAGES** = total remuneration costs; **A_IT_W_AMT** = payments withheld for invoices where no ABN is quoted; **PAYG_TAX_WITHHELD** = PAYG tax; **P_WRK_AMT** = amount withheld from salary, wages and other payments; **T_IT_W_AMT** = amount withheld from investment distributions where no TFN is quoted; **I_INCM_AMT** = PAYG instalment income; **PAYG_INSTMT** = PAYG income tax instalment; **fte** = full time equivalent employees; **hcnt** = headcount of employees; **bg** = business group id.

Source: Suresh et al. (2019)

Box 1 provides some information on common imputed variables and new imputation methods in BLADE.

Box 1 Imputation in BLADE

Researchers typically impute turnover or full time equivalent workers in BLADE when calculating labour productivity. Imputations have typically been fairly simplistic log-log linear regressions to compensate for long tails in both predictors and outcomes. For example, full time equivalents is often defined as:

$$\log(full\ time\ equivalents) = \alpha + \beta \log(wages) + \epsilon$$

While this approximation is a reasonably good fit of the data it restricts use of negative values and could be improved. Newer research has tested various machine learning methods to predict both turnover and full time equivalents. This research found simpler models tended to overfit in a large dataset like BLADE but ensemble method with multiple sub-sampling could produce small errors and good fit. The following shows the mean average of 10 runs of each turnover prediction algorithm using a 90/10 train/test approach:

Algorithm	*No. Features*	*MAE*	*RMSE*	*sMAPE*	*MSE*	R^2	*Time (s)*
Linear Regression	14	0.253	0.381	4.62%	0.145	70.82%	333
Decision Tree	14	0.071	0.236	1.39%	0.056	88.79%	2 003
Ridge Regression	14	0.253	0.381	4.62%	0.145	70.82%	58
Bayesian Ridge	14	0.253	0.381	4.62%	0.145	70.82%	416
LassoCV	14	0.253	0.381	4.62%	0.145	70.82%	1 407
OMPursuitCV	14	0.262	0.392	4.79%	0.154	69.05%	672
Bagging	14	0.06	0.177	1.16%	0.031	93.69%	18 348
Extra Trees	14	0.063	0.174	1.21%	0.03	93.90%	5 709
Gradient Boosting	14	0.074	0.191	1.41%	0.037	92.63%	16 725
Random Forest	14	0.06	0.177	1.16%	0.031	93.70%	17 527
MLP	14	0.078	0.185	1.48%	0.034	93.35%	85 805
GAM	14	0.134	0.244	2.47%	0.06	87.98%	9 472

[a] **MAE** = Mean Absolute Error; **RMSE** = Root Mean Squared Error; **sMAPE** = symmetric Mean Absolute Percentage Error.

Source: Suresh et al. (2019).

Omitted firms

All ABS firm data, not just BLADE, omit some small firms.[5] To some extent, the ABS has no choice in the matter — businesses which earn less than $75 000 in revenue per annum are not required to register an ABN (though there are other incentives to register for these firms).[6] By their nature, businesses without ABNs are not recorded in government administrative data.

[5] See ABS (*Counts of Australian Businesses, including Entries and Exits*, Cat. no. 8165.0) and explanatory notes for the Australian Bureau of Statistics Business Register.

[6] $75 000 represents the threshold for GST liability. A GST liable firm is required to both have an ABN and complete a Business Activity Statement. Low revenue firms may still register an ABN to prevent payers from withholding the top marginal rate of PAYG tax from payments.

The impact of excluding firms without an ABN is not clear as ‘the number of businesses in this category is currently unknown to the ABS’(ABS 2019). Furthermore, the exclusion rule for these firms is a function of size meaning that the population of excluded firms is not the same as the population of included firms. This is particularly important for entry and exit measures as new entrants are likely to start small and may exit before meeting reporting thresholds. Many papers using BLADE thus far have focused on labour productivity using full-time equivalent employment variables. As many of the low revenue firms do not employ (sole traders and similar), exclusion of these firms has not been as problematic.[7]

At this stage, identifying businesses by ABN with over $75 000 revenue in BLADE is acknowledged but accepted as a limitation. There would also be little value in adding firms with less than $75 000 in revenue and an ABN as these firms do not have any reporting liability (e.g. BAS) — i.e. there would be no data for these firms. Given that BLADE will be mostly used to conduct firm level analysis, not something more aggregated, simply identifying the existence of additional firms without any of their characteristics would not add meaningful information to the dataset.

[7] More than three quarters of firms are non-employers in their first year of operation (Bakhtiari 2017).

Appendix A: Who's doing what with BLADE?

Table A.1 **BLADE research overview**

Select current and past BLADE research papers

Organisation	*Topic*	*Citation*	*Description*
Treasury	Productivity Dispersion	(Campbell, Sibelle and Soriano 2019)	Measured the dispersion of labour productivity across firms and across time and found increasing dispersion and evidence of laggards.
ABS/UNSW	Capital measurement and MFP estimation	Unpublished	Attempted to estimate capital stocks and multifactor productivity levels at a firm level. Also attempted to systematise the methodology of dealing with missing data and variables in BLADE.
Treasury	Wage stagnation	(Andrews et al. 2019)	Found that the recent slowdown in wage growth coincided with a breaking of the link between productivity growth and wage growth at a firm level.
Industry	Management capability	(Agarwal et al. 2019)	Industry developed several indices of management capability based on the BCSM, and compared this with similar metrics in the United States and found Australia firms to be lacking.
Industry	Productivity in the manufacturing sector	(Bakhtiari 2019)	Looked at the productivity characteristics of entrants vs incumbents in the manufacturing industry. Noteable as the first attempt to estimate capital stocks and MFP at a firm level in BLADE.
RBA	Credit availability	(Araujo and Hambur 2018)	Using BCS and BIT, looked at the financial characteristics of firms that applied for credit and were either accepted or rejected.
RBA	Concentration and mark up in the retail sector	(Hambur and La Cava 2018)	Using an estimate of the production function, attempted to estimate the mark-ups of retail firms, finding evidence of declining competition in the retail sector.
Industry	Effect of innovation on business growth	(Hendrickson et al. 2018)	Using BCS and BLADE core, found that persistent innovators significantly outgrow their less persistent and non-innovator counterparts in terms of sales, value added, employment and profit growth
Industry	Performance of firms using employee share schemes	(Hendrickson et al. 2017)	Using EAS and BLADE core, found that SMEs were much more likely pay wages in the form of shares in profits. Further those that engaged in share schemes tended to have lower employee churn, higher wages and higher labour productivity.
Industry	Exporter characteristics and performance	(Tuhin and Swanepoel 2017)	Exporters tend to be larger and show superior growth performance prior to exporting, than non-exporters.
Industry	The effect of age on Australian small-to-medium enterprises	(Smith and Hendrickson 2016)	Younger SMEs are more likely to collaborate and engage in innovative behaviour and this is associated with a range of growth variables, including: employment growth, sales growth and productivity and profitability.

Appendix B: BLADE core data item list

Frame (information from the ABS business register)

- BLADE Unit ID
- BLADE Enterprise Group ID
- Time series ID
- State
- Postcode
- Industry Division 2006
- Australian and New Zealand Standard Industrial Classification (ANZSIC), 2006
- Standard Institutional Sector Classification of Australia (SISCA) 2008
- Type of Legal Organisation (TOLO)
- Birth Date (year)
- States of operation
- Alive status
- Non Profit Institution flag
- Private/Public indicator

Business Activity Statement (BAS)

- Total sales
- Export sales
- Other GST-free sales
- Capital purchases
- Non-capital purchases
- Total salary, wages and other payments
- Amount withheld from salary, wages and other payments
- Amount withheld from payment of invoices where no ABN is quoted
- Amount withheld from investment distributions where no TFN is quoted
- Pay as you go (PAYG) tax withheld
- Pay as you go (PAYG) income tax instalment
- Pay as you go (PAYG) instalment income
- Goods and services tax (GST) on sales or GST instalment
- Goods and services tax (GST) on purchases
- Imported goods with GST deferred

Pay As You Go personal income tax

- Head count
- Full time equivalent workers

Business Income Tax

Table B.1 **Variables in Business Income Tax dataset**

Variables included by type of business entity

	Company form	*Partnership form*	*Trust form*	*Individual form*
Business income				
Other sales of goods and services	✓			
Government industry payments assessable for tax (non-primary production)		✓	✓	✓
Government industry payments assessable for tax (primary production)		✓	✓	✓
Total business income/non-primary production		✓	✓	✓
Total business income/primary production		✓	✓	✓
Total business income	✓			✓
Gross distribution from partnerships	✓			
Gross distribution from trusts	✓			
Gross payments where ABN not quoted	✓			
Gross Payments Where Australian Business Number Not Quoted - Non-Primary Production				✓
Gross Payments Where Australian Business Number Not Quoted - Primary Production				✓
Gross rent and other leasing and hiring income	✓			
Other gross income	✓			
Gross Payments – Labour Hire Or Other Specified Payments - Non-Primary Production				✓
Business Income/Gross Pmt - Voluntary Agreement - Pp				✓
Business Income/Gross Pmt - Voluntary Agreement - Npp				✓
Gross Payments – Labour Hire Or Other Specified Payments - Primary Production				✓
Assessable government industry payments	✓			
Gross payments subject to foreign resident withholding	✓			
Income from financial arrangements (TOFA)	✓			
Unrealised gains on revaluation of assets to fair value	✓			
Total dividends	✓			
Forestry managed investment scheme income	✓			
Fringe benefit employee contributions	✓	✓	✓	
Gross interest	✓	✓	✓	
Gross payments subject to foreign resident withholding - NPP		✓	✓	✓
Gross payments subject to foreign resident withholding - PP		✓	✓	✓
Gross payments where ABN not quoted - NPP		✓	✓	
Gross payments where ABN not quoted - PP		✓	✓	

(continued next page)

Table B.1 (continued)

Variables included by type of business entity

	Company form	*Partnership form*	*Trust form*	*Individual form*
Business income				
Other business income - NPP		✓	✓	✓
Other business income - PP		✓	✓	✓
Business expenses				
Cost of sales	✓	✓	✓	✓
Contractor, sub-contractor and commission expenses	✓	✓	✓	✓
Employee superannuation expenses	✓	✓	✓	✓
Bad debts	✓	✓	✓	✓
Lease expenses overseas	✓			
Lease expenses within Australia	✓			
Lease expenses		✓	✓	✓
Rent expenses	✓	✓	✓	✓
Interest expenses overseas	✓	✓	✓	✓
Interest expenses within Australia	✓			✓
Interest incurred on money borrowed from Australian or overseas sources		✓	✓	
Depreciation expenses	✓	✓	✓	✓
Extraordinary items	✓			
Total expenses	✓	✓	✓	✓
Operating profit or loss	✓			
Royalty expenses within Australia	✓			
Royalty expenses overseas	✓	✓	✓	
Total royalty expenses		✓	✓	
Expense reconciliation adjustments		✓	✓	✓
Income reconciliation adjustments		✓	✓	✓
Unrealised losses on revaluation of assets to fair value	✓			
Expenses from financial arrangements (TOFA)	✓			
All other expenses	✓	✓	✓	✓
Motor vehicle expenses	✓	✓	✓	✓
Repairs and maintenance	✓	✓	✓	✓
Foreign resident withholding expenses	✓	✓	✓	✓
Business income or loss				
Total profit or loss	✓			
Net business income or loss (non-primary production)		✓	✓	
Net business income or loss (primary production)		✓	✓	✓
Net business income or loss this year (non-primary production)				✓
Net business income or loss this year (primary production)				✓

(continued)

Table B.1 (continued)
Variables included by type of business entity

	Company form	*Partnership form*	*Trust form*	*Individual form*
Partnerships and trusts				
Distribution from partnerships - Primary production		✓	✓	
Share of net income from trusts - Primary Production		✓	✓	
Deductions relating to distribution in labels A and Z - Primary Production		✓	✓	
Distribution from partnerships, less foreign income - Net-Primary Production			✓	
Distribution from trusts, less net capital gain and foreign income - Non-Primary Production			✓	
Deductions relating to distribution in labels B and R - Non-Primary Production			✓	
Franked distributions from trusts		✓	✓	
Deductions relating to franked distributions from trusts in label F		✓	✓	
Taxable/net income or loss				
Capital works deductions	✓	✓	✓	
Taxable income or loss	✓			
Deferred non-commercial business losses from a prior year - NPP				✓
Deferred non-commercial business losses from a prior year - PP				✓
Total of items 5 to 14 (previously items 4 to 11) on the Partnerships and Trusts tax returns		✓	✓	
Australian franking credits from a New Zealand company	✓			
Deduction for decline in value of depreciating assets	✓			
Deduction for environmental protection expenses	✓			
Exempt income	✓			
Forestry managed investment scheme deductions	✓			
Franking credits	✓			
Immediate deduction for capital expenditure	✓			
Net capital gain	✓		✓	
Non-deductible exempt income expenditure	✓			
Non-deductible expenses	✓			
Offshore banking unit adjustment	✓			
Other assessable income	✓			
Other deductible expenses	✓			
Other income not included in assessable income	✓			
Section 46FA deductions for flow-on dividends	✓			
TOFA deductions from financial arrangements not included in item 6	✓			
TOFA income from financial arrangements not included in item 6	✓			

(continued)

Table B.1 (continued)

Variables included by type of business entity

	Company form	*Partnership form*	*Trust form*	*Individual form*
Taxable/net income or loss				
Tax losses deducted	✓			
Tax losses transferred in	✓			
Deduction for project pool	✓	✓	✓	✓
Landcare operations and business deduction for decline in value of water facility				✓
Landcare operations and deduction for decline in value of water facility	✓		✓	
Section 40-880 deduction	✓	✓	✓	✓
Net income or loss from business this year - Non-primary Production				✓
Net tax assessed	✓			
Taxable or net income	✓			
Amount due or refundable	✓			
Eligible credits	✓			
Franking deficit tax offset	✓			
Gross tax	✓			
Non-refundable carry forward tax offsets	✓			
Non-refundable non-carry forward tax offsets	✓			
PAYG instalments raised	✓			
Refundable tax offsets	✓			
Remainder of refundable tax offsets	✓			
Section 102AAM interest charge	✓			
Subtotal 1	✓			
Subtotal 2	✓			
Subtotal 3	✓			
Tax on taxable income	✓			
Tax payable	✓			
Financial and other information				
Assets (current)	✓	✓	✓	
Total Assets	✓	✓	✓	
Liabilities (current)	✓	✓	✓	
Liabilities	✓	✓	✓	
Inventories (closing)	✓	✓	✓	✓
Purchases and other costs	✓	✓	✓	✓
Inventories (opening)	✓	✓	✓	✓
Trade creditors	✓	✓	✓	✓
Trade debtors	✓	✓	✓	✓
Total salary and wage expenses	✓	✓	✓	✓
Payments to associated persons	✓	✓	✓	✓
Net foreign income	✓	✓		

(continued)

Table B.1 (continued)

Variables included by type of business entity

	Company form	*Partnership form*	*Trust form*	*Individual form*
Financial and other information				
Percentage of foreign shareholding	✓			
Shareholders' funds	✓			
Proprietors' funds		✓	✓	
Attributed foreign income - Listed country	✓			
Attributed foreign income - Section 404 country	✓			
Attributed foreign income - Transferor trust	✓			
Attributed foreign income - Unlisted country	✓			
Commercial debt forgiveness	✓			
Deductions relating to distribution in labels B and R - NPP		✓		
Distribution from partnerships, less foreign income - NPP		✓		
Distribution from trusts, less net capital gain and foreign income - NPP		✓		
Excess franking offsets	✓			
Franked dividends paid	✓			
Franking account balance	✓			
Gross foreign income	✓			
Loans to shareholders and their associates	✓			
TOFA gains from unrealised movements in the value of financial arrangements	✓			
TOFA transitional balancing adjustment	✓			
Tax spared foreign tax credits	✓			
Total TOFA gains	✓		✓	
Total TOFA losses	✓		✓	
Total debt	✓			
Unfranked dividends paid	✓			
Attributed foreign income: FIF/FLP income		✓	✓	
Attributed foreign income: foreign investment fund income	✓			
Attributed foreign income: foreign life policy	✓			
AFI - listed country		✓	✓	
AFI - section 404 country		✓	✓	
AFI - unlisted country		✓	✓	
Foreign income tax offset	✓	✓	✓	
Loss carry-back tax offset	✓			
Tax loss for middle year (not already utilised) chosen to be carried back to earliest year	✓			
Tax loss for current year chosen to be carried back to earliest year	✓			
Net exempt income for earliest year	✓			
Income tax liability for earliest year	✓			

(continued)

Table B.1 (continued)
Variables included by type of business entity

	Company form	*Partnership form*	*Trust form*	*Individual form*
Financial and other information				
Percentage of non-member income	✓			
Unfranked amount		✓	✓	
Franked amount		✓	✓	
Franking credit		✓	✓	
Gross rent		✓	✓	
Other Australian income		✓	✓	
Interest deductions		✓	✓	
Gross - Other assessable foreign source income		✓	✓	
Net - Other assessable foreign source income			✓	
Net rent		✓	✓	
Other rental deductions		✓	✓	
Deductions relating to Australian investment income			✓	
Credit for TFN amounts withheld from payments from closely held trusts			✓	
Deductions relating to Franked distributions			✓	
TFN amounts withheld from dividends			✓	
Early stage investor tax offset - current year tax offset	✓			
Early stage investor tax offset - tax offset carried forward from a previous year	✓			
Early stage venture capital limited partnership tax offset - current year tax offset	✓			
Early stage venture capital limited partnership tax offset - tax offset carried forward from a previous year	✓			
Credit for TFN amounts withheld from payments from closely held trusts			✓	
Total amount of deductions against PSI included at item 5 expense labels			✓	
Share of franking credits from franked distributions			✓	
Forestry management investment scheme deduction			✓	
Forestry management investment scheme income			✓	
Share of credit for tax withheld - foreign resident withholding (excluding capital gains)			✓	
Total amount of PSI included at item 5 income labels			✓	
Total amount of deductions against PSI included at item 5 expense labels			✓	

(continued)

Table B.1 (continued)

Variables included by type of business entity

	Company form	*Partnership form*	*Trust form*	*Individual form*
Financial and other information				
Total other deductions			✓	
Share of credit for TFN amounts withheld from interest, dividends and unit trust distributions			✓	
Share of credit for tax withheld where ABN not quoted			✓	
Capital allowances				
Other depreciating assets first deducted	✓	✓	✓	✓
Intangible depreciating assets first deducted	✓	✓	✓	✓
Termination value of intangible depreciating assets	✓	✓	✓	✓
Termination value of other depreciating assets	✓	✓	✓	✓
Assessable balancing adjustments on the disposal of intangible depreciating assets		✓	✓	
Deductible balancing adjustments on the disposal of intangible depreciating assets		✓	✓	
Total adjustable values at end of income year		✓	✓	
Research and Development				
Accounting expenditure in item 6 subject to R&D tax incentive	✓			
Australian owned R&D tax concession – not including label M	✓			
Australian owned R&D – extra incremental 50% deduction	✓			
R&D tax offset, if chosen	✓			
Refundable R&D tax offsets	✓			
Non-refundable R&D tax offsets	✓			
R&D recoupment tax	✓			
R&D tax offset	✓			
Non-refundable R&D tax offset carried forward from previous year	✓			
Non-refundable R&D tax offset carried forward to next year	✓			
Non-refundable R&D tax offset to be utilised in current year	✓			

References

ABARES (Department of Agriculture) 2019, *Introducing ABARES farmpredict*, https://www.agriculture.gov.au/abares/research-topics/working-papers/farmpredict (accessed 4 September 2020).

ABS (Australian Bureau of Statistics) 2015, *Australian System of National Accounts Concepts, Sources and Methods*, Cat. No. 5216.0.

—— (Australian Bureau of Statistics) 2019, *Explanatory Notes - Counts of Australian Businesses, including Entries and Exits*, https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/8165.0Explanatory+Notes1June%202014%20to%20June%202018 (accessed 22 January 2020).

—— (Australian Bureau of Statistics and Statistics New Zealand) 2006, *Australian and New Zealand Standard Industrial Classification (ANZSIC)*, Cat. no. 1292.0, Canberra.

Agarwal, R., Bajada, C., Brown, P., Morgan, I. and Balaguer, A. 2019, *Development of Management Capability Scores*, Research Paper, Department of Industry, Innovation and Science.

Andrews, D., Deutscher, N., Hambur, J. and Hansell, D. 2019, *Wage Growth in Australia: Lessons from Longitudinal Microdata*, Treasury.

—— and Hansell, D. 2019, *Productivity-enhancing labour reallocation in Australia*.

Araujo, G. and Hambur, J. 2018, 'Which Firms Get Credit? Evidence from Firm-level Data', *RBA Bulletin*, vol December Quarter.

ATO (Australian Tax Offfice) 2021, *Registering for GST*, Australian Tax Office, https://www.ato.gov.au/Business/GST/Registering-for-GST/#:~:text=You%20must%20register%20for%20GST,the%20first%20year%20of%20operation (accessed 14 April 2021).

Bakhtiari, S. 2017, *Entrepreneurship Dynamics in Australia: Lessons from Micro-data*, Department of Industry, Innovation and Science.

—— 2019, *Do manufacturing entrepreneurs in Australia have (or develop) a productivity advantage?*, 7/2019, Department of Industry, Innovation and Science.

Campbell, S., Sibelle, A. and Soriano, F. 2019, *Measuring Productivity Dispersion in Selected Australian Industries*, Treasury-ABS Working Paper, Treasury.

Foster, L., Haltiwanger, J. and Syverson, C. 2008, 'Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?', *American Economic Association*, vol. 98, no. 1, pp. 394–425.

Fox, K. 2019, 'MFP Measurement Using BLADE: Insights, Challenges and Future Directions', https://www.business.unsw.edu.au/Campaigns-Site/emg-workshop-2019/Documents/Fox_MFP_BLADE_EMG_5Dec2019.pdf (accessed 18 December 2019).

Hambur, J. and La Cava, G. 2018, 'Business Concentration and Mark-ups in the Retail Trade Sector', *RBA Bulletin*, vol December Quarter`.

Hansell, D., Nguyen, T. and Soriano, F. 2015, *Can we improve on a headcount? Estimating unobserved labour input with individual wage data*, Paper presented to the Australian Labour Market Research Workshop, Adelaide December.

Hendrickson, L., Pachernegg, T., Boyle, M., Bucifal, S. and Hansell, D. 2017, *The performance and characteristics of Australian firms with Emkployee Share Schemes*, Research Paper.

——, Taylor, D., Ang, L., Cao, K., Nguyen, T. and Soriano, F. 2018, *The impact of persistent innovation on business growth*, Research Paper.

Little, R. and Rubin, D. 2019, *Statistical Analysis with Missing Data*, 3rd edn, John Wiley & Sons, Inc.

Smith, R. and Hendrickson 2016, *The effect of age on Australian small-to-medium enterprises*, Research Paper, Department of Industry, Innovation and Science.

Suresh, M., Taib, R., Zhao, Y. and Jin, W. 2019, ‘Sharpening the BLADE: Missing Data Imputation Using Supervised Machine Learning’, Liu, J. and Bailey, J. (eds), presented at the *AI 2019: Advances ing Artificial Intelligence*, Springer International Publishing, Cham, pp. 215–227.

Tuhin, R. and Swanepoel, J.A. 2017, *Export behaviour and business performance: Evidence from Australian microdata*, Research Paper.

United States Census Bureau 2021, *Longitudinal Business Database*, United States Census Bureau, https://www.census.gov/programs-surveys/ces/data/restricted-use-data/longitudinal-business-database.html (accessed 14 April 2021).