Ebrahim, Amina; Kreuser, Friedrich; Kilumelume, Michael

Working Paper

The guide to the CIT-IRP5 panel version 4.0

WIDER Working Paper, No. 2021/173

Provided in Cooperation with:
United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: Ebrahim, Amina; Kreuser, Friedrich; Kilumelume, Michael (2021) : The guide to the CIT-IRP5 panel version 4.0, WIDER Working Paper, No. 2021/173, ISBN 978-92-9267-113-6, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki,
https://doi.org/10.35188/UNU-WIDER/2021/113-6

This Version is available at:
https://hdl.handle.net/10419/249479

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2021/173

The guide to the CIT-IRP5 panel version 4.0

Amina Ebrahim,[1,*] C. Friedrich Kreuser,[2] and Michael Kilumelume[1,3]

November 2021

United Nations University World Institute for Development Economics Research
wider.unu.edu

Abstract: This paper presents version 4.0 of the CIT-IRP5 firm-level panel dataset. Version 4.0 is the latest edition of the firm-level component of the combined administrative data using sources from the South African Revenue Service. We show that differences in forms and vintages do generally not preclude consistent identification of output, employment, cost of sales, and capital stock over time. We discuss the inclusion of contributions by other researchers in the classification of multi-national firms, industry, and employment income. The tax dataset is shown to have within-firm consistency with previous versions of the data, and the improvements to the data result in greater consistency with external data sources.

Key words: administrative data, tax, firm-level

JEL classification: C55, C80

Acknowledgements: The task to create the CIT-IRP5 panel would not have been possible without the meticulous work by Daniel Brink, Singita Rikhotso, Junior Chiweza, Michelle Pleace, Mlungisi Ndlovu, and Grace Bridgman. We also acknowledge the feedback and support from Aalia Cassim, Marlies Piek, Andrew Nell, and Jacobus Nel. We thank Carol Newman, John Rand, Jukka Pirttilä, Murray Leibbrandt, Catherine MacLeod, and participants of the SA-TIED Metadata work-in-progress meeting (December 2019) for assistance, comments, and suggestions at various stages of this project. Any errors, of course, remain our own. C. Friedrich Kreuser gratefully acknowledges financial support from the United Nations University World Institute for Development Economics Research for this project.

[1] UNU-WIDER, Helsinki, Finland; [2] Trinity College Dublin, Dublin, Ireland; [3] Stellenbosch University, Stellenbosch, South Africa; * corresponding author: Amina Ebrahim, amina.ebrahim@wider.unu.edu

This study has been prepared within the UNU-WIDER project Southern Africa—Towards Inclusive Economic Development (SA-TIED).

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9267-113-6

https://doi.org/10.35188/UNU-WIDER/2021/113-6

Typescript prepared by Ayesha Chari.

United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

This paper describes version 4.0 of CIT-IRP5 panel and is designed to provide some guidance to researchers using the data (including recommended citations, see Appendix A). The unbalanced firm-level panel was created in June 2016 (version 1.0), updated at the end of 2016 (version 2.0) and updated again in February 2019 (version 3.0) with the latest tax data available.[1] The panel was created through collaboration with the National Treasury, the South African Revenue Service (SARS), and the United Nations University World Institute for Development Economics Research (UNU-WIDER).

Although this paper follows the approach of Pieterse et al. (2018) in the creation of the panel, it can be read as a standalone document. We highlight the qualitative changes made to the data, including changes in the construction process, the implication of changes in the vintages, improvements to employment aggregation, updated industry variables, as well as changes in the capital and cost of sales variables. We further show the consistency of the dataset by comparing the updated aggregates to those of Statistics South Africa (Stats SA) and computing the within-firm between-version differences of key variables.

The paper is organized as follows. Section 2 introduces the panel dataset and highlights the main components of the data along with its main unit of account and underlying sources. Section 3 discusses the qualitative changes in the data. Section 4 details the quantitative changes and improvements made to aggregation and assignment. Section 5 compares the data with external and previous datasets. Section 6 concludes.

2 The CIT-IRP5 firm-level panel

The firm panel harmonizes business-level corporate income tax, job-level employee income tax, VAT entity-level value-added tax (VAT), and transaction-level customs tax data into a single firm-level dataset. A firm, for purposes of the panel, refers to a business resident that may or may not include several different branches, plants, and subsidiaries.[2] This section briefly discusses these underlying data sources for new users, and identifies the underlying forms, unit of analysis, and cross data matching.

2.1 Firm-level data

The corporate income tax data include data on total sales, cost of sales, capital stock, employee costs, and tax allowances, among others. The corporate tax data (CIT) represent the universe of corporate tax returns submitted to SARS on or before the end of January 2020 for the tax years 2008–2018. Firms are identified by an anonymized tax reference number, which is treated as the unique identifier for each entity. We refer to the income statement and balance sheet data of firms as the CIT data.

[1] Versioning of the panel was strictly applied from 2019 onwards.

[2] The present version of the data does not incorporate detailed information on corporate structure of firms, which is required by Taxation Laws Amendment Act 23 of 2018. When referring to the Income Tax Act, or simply the Act, we refer to Income Tax Act 58 of 1962 (see South Africa 1962).

A tax-registered entity is required to submit its tax return within 12 months of its financial year end, which will always be treated as the entity's tax year from the perspective of the tax form (SARS 2020a). A firm with a financial year end in December 2017 is thus not required to submit returns until December of 2018. The latest data used for this update were obtained from SARS from November 2019 and January 2020, meaning that the 2018 financial year will likely be incomplete due to reporting. It should be noted that this lag is persistent, with 2017 still missing several returns.

The CIT data available to researchers are the result of harmonizing corporate tax returns over time. Corporate tax returns are submitted to SARS via the IT14 and ITR14 forms. The ITR14 form replaced the IT14 on 4 May 2013, requiring harmonization of the different fields.[3] The initial ITR14 form has undergone minor changes over time. Where the change in the form affects sales, cost of sales, fixed or tangible capital stock, or employment, the changes in the forms are harmonized to ensure that variables consistently represent the underlying economic information over time.

An example of a major change in the CIT form is the introduction of the 'Vehicle amount' field from 2016 onwards. The inclusion of this field necessitated the adjustments to the panel's aggregation of capital stock variables, as discussed in Section 4.1.1. An example of an unadjusted change in the underlying ITR14 data relates to the firm's consolidation status. The decrease in the proportion of companies responding 'yes' to being 'part of a group of companies that prepare consolidated financial statements' in the 2017 is a 'ctax' year that results from changes to the ITR14 form in 2017. Before 2017, the same field asked the taxpayer to declare whether the company is a subsidiary of a group of companies as defined in Section 1 of the Income Tax Act. According to Section 1 of the Income Tax Act, a group has at least 70 per cent ownership. On the other hand, the accounting consolidation implies a more than 50 per cent ownership or control criterion. Therefore, the ITR14 form version change in 2017 has a selection effect for the variable associated with this field. Researchers should also be aware that some variables in the dataset suffer from selection bias resulting from changes in reporting requirements associated with the switch from the IT14 to the ITR14 form. For instance, while all company types reported the research and development expenditure amounts on the IT14 form, only medium to large firms do so on the ITR14 form. As a result, in the post-2012 period, the research and development variable (*x_rd*) only contains values for medium to large firms. The same is true for other variables affected by the change in reporting requirements. Owing to the size of the data and the sheer number of questions, we cannot ensure that we are aware of every single one of the variables with underlying content or context changes; researchers should thus always refer to the SARS user guides for the relevant years when using the data (SARS 2020a, 2020b).

The present version of the data includes IT14 forms from 2008 to 2013 and ITR14 forms from 2008 to 2018. The CIT data available to researchers in the present version do not attempt to harmonize data across vintages, meaning that data that were available in the raw data used in previous versions but that became unavailable in the latest raw data will be missing. Vintage issues and potential implications are discussed in Section 3.6.

2.2 Employee tax data

The employee tax data contain job-level data that include information on the date of birth and income sources for persons employed in a pay-as-you-earn (PAYE) registered entity. The data reflect the universe of IRP5 and IT3(a) certificates submitted to SARS from 2008 to 2019. SARS

[3] Details of the field overlaps are available on request.

sets the deadline for employer reconciliation on the last day of May of the relevant tax year. That is, the forms for workers with employment periods ending on the last day of February of the given year must be submitted by the last day of May of the same year (SARS 2020e). Each job is identified by a unique form number, which belongs to a person identified by a unique identification number or passport number that is linked to a PAYE reference number.[4] PAYE reference numbers are linked to the unique firm identifier—the tax reference number—of the CIT data using a correspondence table provided by SARS. A tax reference number is sometimes linked to several PAYE numbers. The employee data are aggregated to the firm level based on the CIT tax reference number (*taxrefno*), as discussed in Section 4.2. This has implications for matching the employee data with the CIT data, which we discuss in Section 5.1. We discuss the employment data and their vintages in Section 3.6.

2.3 International trade data

Data on international trade are aggregated from transaction-level data recorded through customs declaration forms. The declaration form—SAD 500—requires customs registered entities to submit information on the Rand value, HS6 product code, and partner country at both port of trade and final destination/origin.

The present trade data reflect only the trade component of the customs data received and removes entries related to the warehousing of reported goods.[5] We assign each transaction made to the relevant firm's financial year end through the date information provided in the CIT forms. The present version of the panel expands the available customs fields by including information on the total trade with specific regional trading blocks,[6] specific levels of income, the number of different HS6 and HS4 product lines exported and imported, as well as information on the largest regional partners and amounts.

2.4 VAT data

The VAT data are collected from the VAT101, VAT102, and VAT201 forms submitted to SARS by the registered VAT vendors. The registration for VAT is compulsory for entities with taxable supplies (i.e., sales or turnover exceeding R1 million in any 12 months) and voluntary for taxable supplies less than R1 million.[7] The VAT201 forms constitute most information available in the VAT data, with only industry information from the VAT101 and VAT102 tax forms.

The VAT data make up a transactional-level dataset in that they record a transaction between the VAT filing entity and the revenue service; they do not include identifiable information on individual transactions between firms where a counterparty can be identified.[8] VAT forms can reflect information at the monthly, bi-monthly, annual, and bi-annual level (SARS 2020d). The

[4] We only observe anonymized versions of these identifiers.

[5] These goods were identified under consultation from SARS. Documentation regarding this process is available on request.

[6] These blocks include East Asia and Pacific, Europe and Central Asia, Middle East and North Africa, North America, South Asia, and Sub-Saharan Africa as well as the World Bank's classification of high-income non-OECD, high-income OECD, low income, lower middle income, and upper middle income according to the latest classification (World Bank 2020). We also include a separate indicator for trade with members of the Southern African Development Community.

[7] See Chapter 6 of SARS (2019) for details on taxable supplies.

[8] This information is not excluded; it does not exist.

VAT data are aggregated to the corresponding firms' annual level before the transactions are allocated to the corresponding firm in a given tax year. The VAT variables reported in the CIT-IRP5 panel, therefore, correspond to the same underlying period as the information contained in the CIT data. The VAT data are linked to the CIT data using the correspondence tables provided by SARS.

The VAT panel is available as a separate firm-year-level dataset and VAT record level dataset in addition to its inclusion in the CIT-IRP5 panel. Variables from the VAT data include the prefix '*v_*' in the CIT-IRP5 panel. Owing to limited changes in the VAT data, we do not discuss its aggregation in a specific section. Researchers should note that the VAT rate was increased from 14 to 15 per cent in April 2018 after being announced in the February 2018 budget speech (National Treasury 2018).

3 Qualitative changes

This section highlights specific qualitative changes made to the data available to researchers. Key improvements in the present version of the data are the availability of a multinational firm identifier, a comprehensive industry identifier, and improved employment identifiers. Certain changes in the underlying data required a different approach to the aggregation of data and availability of certain variables.

3.1 Multinationals

CIT-IRP5 version 4.0 includes a set of variables related to identifying firms with international connections. Kilumelume et al. (2021) fully describe the classification of firms according to a strict foreign indicator, a broad foreign indicator, a domestic or foreign multinational indicator, and a foreign connection indicator.

Firms are considered strictly foreign (*ITR14_c_foreign_strict*) where they state that their ultimate holding company is resident outside South Africa. Firms are considered broadly foreign (*ITR14_c_foreign_broad*) if (i) they state that they are not a resident in South Africa for income tax purposes, or (ii) they are resident outside South Africa by virtue of a treaty to avoid double taxation, or (iii) the return submitted is with respect to a branch, permanent establishment, or agency of a foreign company, or (iv) the firm reports total dividends subject to double taxation relief.

Firms are considered a multinational with a foreign parent (*ITR14_c_mne_ type*) according to the strict definition above. Firms are considered a domestic parent of a multinational firm if the ultimate holding company is not a company resident outside South Africa and if any of the following conditions are true: (i) the company is part of a multinational enterprise, (ii) the company has claimed an exemption to foreign dividends according to s10(1)(k)ii(dd) or s10B(2)(a) of the Income Tax Act, (iii) any foreign dividend was subject to participation exemption, or (iv) the company directly or indirectly controls more than 10 per cent of the total participation rights or voting rights in a Controlled Foreign Company (s9D of the Income Tax Act; see South Africa 1962).

The foreign connection indicator (*ITR14_c_fcf*) includes firms that (i) are headquarter companies with minimal asset rules, where at least 80 per cent or more of the cost of the total assets are attributable to a qualifying foreign company, or (ii) have participation or voting rights in a controlled foreign company, or (iii) have foreign dividends exempt in terms of s10B(2)(a) of the

Income Tax Act, or (iv) have foreign dividends subject to participation exemption, or (v) have foreign income and expenditure in terms of s31(1)(a) in the Income Tax Act (South Africa 1962).

3.2 Industry data

Firm-level industry classifications are significantly improved from previous versions. The present dataset uses the industry classification developed by Budlender and Ebrahim (2020). The classification is based on the Main Industry Code submitted in the ITR14 or IT14 forms. Missing values are iteratively imputed using within-firm time-neighbouring values.[9] Budlender and Ebrahim (2020) show that the imputed code improves significantly on the raw code as it is internally consistent and broadly matches industry classification in other sources of South African data. The SIC 7 is based on the ISIC Revision 4 classification system but adjusted for the South African economic environment (Stats SA 2012). The remainder of this paper uses this measure of industry whenever industry is used. Researchers wishing to use only the IRP5 data are referred to Budlender and Ebrahim (2020) for detailed information about industry classification in the payroll data.

3.3 Deflators

The economy-wide and industry-specific deflator variables in the CIT-IRP5 version 4.0 are based on information by the South African Reserve Bank (SARB). The deflator series is obtained from the SARB Quarterly Bulletin National Accounts (SARB 2020a) and Business Cycle (SARB 2020b) data, and includes (i) gross fixed capital formation, (ii) gross domestic product, (iii) producer price index, (iv) consumer price index, and (v) gross value added. Industry-specific deflators include (i) producer price index, (ii) consumer price index, (iii) gross value added, and (iv) gross fixed capital formation. Industry-level deflators are merged into the panel using the imputed two-digit main industry code in SIC 5 format (Budlender and Ebrahim 2020). The industry datasets (*citirp5_industry_v4.dta*), constructed by Budlender and Ebrahim (2020), and deflator datasets (*deflators_v4.dta*) are available to researchers in the secure data facility. Researchers have the option of merging in the deflators using their preferred industry classification variable.

3.4 Employment data

Previous versions of the panel provided several measures of employment, based on weighting, employment measure, and nature of person status. The main qualitative addition to the employment measure is the identification of labour income for employees based on income codes identified by Kerr (2020). Kerr (2020) shows that these measures are more appropriate by measuring his selection of income code against previous methods that used other income sources codes.

Substantial improvements have been made to the weighting of employees over time based on both the number of days worked and the number of periods worked, as discussed in Section 4.2.[10] When referring to employees we use Kerr's (2020) measure of employment weighted by the periods data using the basic nature of person data. When comparing the employment data with previous versions, we use the Pieterse et al. (2018) measure of employment.

---

[9] This field is relatively incomplete before 2013.

[10] PAYE paying entities report a start date and end date as well as a 'periods worked' and 'total periods' measure. We discuss these measures in Section 4.2.2.

3.5 Temporal matches and financial year

As discussed in Section 2, a firm's financial year end determines the income statement and balance sheet information reported to the revenue service. The employment data, VAT, and customs data, on the other hand, is intended to always start on 1 March of the previous year and end on the last day of February of the tax year in question. This means that we cannot simply link transactions or employees to firms based only on the tax year information, as the periods in question will not overlap one to one. We harmonize the different data sources by aggregating information from the VAT, customs, and IRP5 datasets to the relevant firm year end. We discuss the aggregation procedure in Section 4. The aggregation procedure implies that the closing date connected to the tax year of the information in the CIT component of the data forms the unique temporal identifier per firm.

Where the year information is used to identify year effects, the differences in the financial year end of firms add noise to the effect. A firm ending in December will be grouped with a firm ending in January of the same year, instead of being grouped with a firm that ends in February of the next year. In previous versions of the released data, this mismatch was corrected by creating a firm year variable equal to the tax year only when a firm's financial year end is before August. Firms with financial year end months after July are assigned a financial year that is equal to the following tax year. The previous version of the data manipulated the field to ensure that it could be used to specify the data as a panel with no temporal overlaps, thereby smoothing over changes in a firm's financial year end by assigning the financial year end month as the mode of reported end month. The present version of the data no longer makes this assumption in the construction of the financial year end variable and the researcher is tasked with constructing the appropriate temporal fixed effect for their purpose.

In Table 1, we show how firms reporting changes in their financial year end around the cut-off month will be assigned to differing financial year ends. These changes result in a situation where a firm's financial year can be missing, or the same financial year can be assigned to two different tax years. The present approach reduces the risk of researchers using a non-detailed temporal fixed effect.

Table 1: Tax year and financial year matching

Tax reference number	Tax year	Month of financial year end	*finyear* variable	Match
AA	2008	12	2009	—
AA	2009	12	2010	—
AA	2010	11	2011	—
AA	2011	11	2012	—
AB	2008	6	2008	—
AB	2009	6	2009	Missing *finyear* (2010)
AB	2010	8	2011	Missing *finyear* (2010)
AB	2011	8	2012	—
AC	2008	10	2009	—
AC	2009	10	2010	Repeated *finyear* (2010)
AC	2010	6	2010	Repeated *finyear* (2010)
AC	2011	6	2011	—

Note: this table shows the 'Missing *finyear'* and 'Repeated *finyear'* issues associated with the *finyear* variable in CIT-IRP5 v4.0. '—' indicates that the *finyear* variable is non-missing nor repeating.

Source: authors' illustration, based on data in v4.0 of the CIT-IRP5 panel (National Treasury and UNU-WIDER 2021a).

3.6 Vintages

As the administrative database underlying the CIT-IRP5 matures older datasets will no longer be extracted by the revenue service, meaning that the information for certain years is based on information that may be in a different shape, form, or level of completion. We refer to various extractions of the data as vintages. Different vintages do not always come with their own correspondence tables. The correspondence tables allow for the linking of firms across tax forms and are thus crucial in identifying employment, customs, and VAT information in firms. At present, the correspondence table received by SARS is not inclusive of previous versions, meaning that certain entities existing in previous periods are not always included in the latest correspondence table. The exact reason for this has not been identified. At present, no direct documentation could be found on the re-use and recycling of PAYE, VAT, or customs reference numbers. In this context, we use the correspondence table relevant to the specific vintage and incorporate the correspondence information from other sources using a nearest temporal neighbour approach with a preference for preceding data in the case of ties.

In Table 2, we provide the underlying extraction for each component of CIT-IRP5 version 4.0. The IT14 and ITR14 data are based on the same data as version 3.0 from 2008 to 2011. The remainder of the databases are updated considerably. The implication of differing vintages and extractions are varied. Certain variables will not be available in certain extractions, this may be due to changes in the extraction procedure or changes in the form, researchers using the introduction of a specific tax or changes in form should take care to ensure they understand exactly how the change corresponds to changes in source before making causal inferences. Different extraction may also implicitly collapse certain fields, we discuss capital stock as an example in Section 4.1.1. While we take every step to ensure harmonization, changes of this nature require updating the underlying cleaning procedures and may also result in differing totals both within and between firms. At present, SARS and the National Treasury are working to harmonize the extraction process to ensure that issues of this nature are limited. Researchers should further note the change in transfer pricing and capital gains data.

Table 2: Vintages and underlying sources

Tax year	IT14		ITR14		Transfer pricing and capital gains		VAT		IRP5		Customs	
2008	Jul-15	A	Jul-15	A	NA	NA	—		Aug-17	D	—	—
2009	Jul-15	A	Jul-15	A	NA	NA	Jun-19	C	Aug-17	D	Jun-19	C
2010	Jul-15	A	Nov-19	B	Jul-17	—	Jun-19	C	Aug-17	D	Jun-19	C
2011	Jul-15	A	Nov-19	B	Jul-17	—	Jun-19	C	Aug-17	D	Jun-19	C
2012	Jul-15	A	Nov-19	B	Nov-19	B	Jun-19	C	Aug-17	D	Jun-19	C
2013	Jul-15	A	Nov-19	B	Nov-19	B	Jun-19	C	Aug-17	D	Jun-19	C
2014	NA	NA	Nov-19	B	Nov-19	B	Jun-19	C	Aug-17	D	Jun-19	C
2015	NA	NA	Nov-19	B	Nov-19	B	Jun-19	C	Oct-19B	B*	Jun-19	C
2016	NA	NA	Nov-19	B	Nov-19	B	Oct-19	B	Oct-19B	B*	Nov-19	B
2017	NA	NA	Nov-19	B	Nov-19	B	Oct-19	B	Oct-19B	B*	Nov-19	B
2018	NA	NA	Nov-19	B	Nov-19	B	Oct-19	B	Oct-19B	B*	Nov-19	B
2019	NA	NA	—	—	—	—	—	—	Oct-19B	B*	—	—

Note: the table describes which dataset was used and the related correspondence table from different vintages. Oct-19B refers to an extraction different from the October 2019 extraction. B* is the correspondence table B but did not come with the October 2019 extraction. 'NA' indicates that there is no specific data from SARS containing this information. '—' indicates that the data exist but are not available at the time of writing.

Source: authors' construction.

The identification of a firm is only possible where the firm's tax reference number is linked to a single *partyid*, the field that links tax reference numbers to PAYE, VAT, and customs reference numbers. Previous versions simply ignored cases where a single source (PAYE, VAT, or customs) reference number is linked to multiple tax reference numbers. The introduction of the *partyid* field allows for multiple *taxreference* numbers to be linked to the same entity. These cases do not allow for IRP5, VAT, or customs data to be incorporated as we cannot assign information from these sources to *taxreference* numbers consistently. We do not assign these firm employee data as we cannot ensure the quality of match. In Table 3, we show that only around 0.5 per cent of firms have these conflicting records. While this number is low, the sales aggregates of firms with these conflicting records accounts for about 13 per cent of total sales. At the time of writing this issue is still under review.

Table 3: Observations and sales by multiple record status

	2009	2010	2011	2012	2013	2014	2015	2016	2017
Observations	595,980	754,750	757,522	813,419	833,077	855,929	873,371	918,050	908,405
No multiple records	593,192	751,372	754,214	810,098	829,903	853,027	870,586	915,400	905,904
	99.53%	99.55%	99.56%	99.59%	99.62%	99.66%	99.68%	99.71%	99.72%
Multiple records	2,788	3,378	3,308	3,321	3,174	2,902	2,785	2,650	2,501
	0.47%	0.45%	0.44%	0.41%	0.38%	0.34%	0.32%	0.29%	0.28%
Sales	4,442,200	6,128,060	6,450,590	7,249,986	7,840,760	8,528,698	8,906,206	9,402,557	9,837,994
Observations	316,761	329,549	327,451	333,924	340,198	344,884	349,952	358,410	356,966
No multiple records	4,202,827	5,424,465	5,721,483	6,298,131	6,793,238	7,397,236	7,695,357	8,156,925	8,566,208
	94.61%	88.52%	88.70%	86.87%	86.64%	86.73%	86.40%	86.75%	87.07%
Observations	314,478	327120	325,095	331,548	337,880	342,634	347,777	356,337	355,024
	99.28%	99.26%	99.28%	99.29%	99.32%	99.35%	99.38%	99.42%	99.46%
Multiple records	239,373	703,596	729,107	951,855	1,047,522	1,131,462	1,210,850	1,245,632	1,271,786
	5.39%	11.48%	11.30%	13.13%	13.36%	13.27%	13.60%	13.25%	12.93%
Observations	2,283	2,429	2,356	2,376	2,318	2,250	2,175	2,073	1,942
	0.72%	0.74%	0.72%	0.71%	0.68%	0.65%	0.62%	0.58%	0.54%

Note: this table shows the number of observations in the total sample in the first row followed by the number of observations without and with multiple records. The percentage is the proportion of firms satisfying the criteria in all observations. The totals in the Sales section show the unadjusted sum of sales for all firms with the number of non-missing observations. No multiple records and multiple records are the sum of sales for all firms satisfying said criteria in a given year. The observations are the number of firms satisfying the criteria.

Source: authors' calculations based on CIT-IRP5 firm-level panel data v4.0 (National Treasury and UNU-WIDER 2021a).

3.7 Dormant firms

In the present paper, we consider a firm to be dormant[11] where it indicates either that it is dormant or that it became dormant in the year of assessment. In Table 4, we show that around 36 per cent of all CIT forms from 2010 to 2018 belong to dormant firms, with the figure being closer to 42 per cent of firms from 2016 onwards. We also show that a negligible proportion, below 1 per cent, of these firms report sales data. In the remainder of this paper, we will exclude these firms from analysis unless otherwise stated.

Table 4: Dormant firms

Tax year	All firms				With sales		
	Total firms	Dormant	Dormant YOA	Any dormant	Dormant	Dormant YOA	Any dormant
2010	754,750	183,362	7,967	183,362	1,032	138	1,032
		24.29%	1.06%	24.29%	0.31%	0.04%	0.31%
2011	757,522	213,230	12,202	213,230	552	207	552
		28.15%	1.61%	28.15%	0.17%	0.06%	0.17%
2012	813,419	273,665	26,916	273,665	1,122	580	1,122
		33.64%	3.31%	33.64%	0.34%	0.17%	0.34%
2013	833,077	306,092	30,296	306,094	1,935	1,026	1,936
		36.74%	3.64%	36.74%	0.57%	0.30%	0.57%
2014	855,929	332,551	26,383	332,559	1,156	1,162	1,162
		38.85%	3.08%	38.85%	0.34%	0.34%	0.34%
2015	873,371	345,932	25,818	345,934	1,136	1,135	1,137
		39.61%	2.96%	39.61%	0.32%	0.32%	0.32%
2016	918,050	382,754	27,132	382,754	1,420	1,417	1,420
		41.69%	2.96%	41.69%	0.40%	0.40%	0.40%
2017	908,405	381,209	25,577	381,209	1,429	1,427	1,429
		41.96%	2.82%	41.96%	0.40%	0.40%	0.40%
2018	788,527	322,211	20,457	322,211	1,389	1,386	1,389
		40.86%	2.59%	40.86%	0.43%	0.43%	0.43%
Total	7,503,050	2,741,006	202,748	2,741,018	11,171	8,478	11,179
		36.53%	2.70%	36.53%	0.36%	0.28%	0.36%

Note: this table shows the number of by dormancy status in Columns 3–5, and the number of firms with sales by dormancy status in Columns 6–9. The percentage figures are the proportion of firms satisfying the dormancy and data availability criteria in total firms in the year.

Source: authors' calculations based on CIT-IRP5 firm-level panel data v4.0 (National Treasury and UNU-WIDER 2021a).

4 Building CIT-IRP5 version 4

This section discusses the construction of version 4.0 of the data. We highlight the changes made to the aggregation of fields and changes in key variables for the CIT data and then discuss the construction of the employment data. We provide some advice for appropriate employment measures and the construction of balanced books.

[11] In the dataset, the company type variable (*c_type*) classifies a company as dormant if it responds 'yes' to the question 'is the company dormant?' and 'no' to 'did the company become dormant/inactive during the year of assessment?'. Thus, companies that become dormant during the year of assessment are not regarded as dormant companies.

4.1 Corporate income tax data

The corporate income tax data in version 4.0 are based on the underlying data reported in Table 2, which itself comes from the ITR14 and IT14 forms. In the construction of the data, care is taken to keep variable names consistent between versions of the panel. Users should note that this is despite changes in the structure and naming conventions in the underlying raw data.[12] Appendix Table A1 provides a list of the prefixes used in the panel along with their descriptions.

The main harmonization between CIT forms relate to matching the IT14 and ITR14 forms. The present paper follows the same procedure as in Pieterse et al. (2018) in which ITR14 data are used where available and only replaced with IT14 data where not available. The field number of the IT14 data is usually specified in the label of the harmonized variables to allow the user to confirm that the correct underlying data were used.

4.1.1 Aggregating fields and changes in underlying data

The ITR14 tax form is separated into three mutually exclusive firm-category sections, each requiring differing levels of detail in the income statement and balance sheet. In increasing order of detail required these categories are (i) micro business/body corporate/share block companies, (ii) small business and dormant companies, and (iii) medium to large business companies. The underlying raw data in previous versions were disaggregated in this structure and harmonized within the cleaning procedure. SARS harmonized these fields in their extraction process for version 4, meaning that harmonization for several variables is no longer required. It should be noted that SARS do not harmonize all fields. In Table 5, we show aggregation of the property and equipment for three hypothetical firms. In the latest raw data, the property, plant, and equipment fields for micro and small firms is received aggregated into a single field; in previous versions, the raw data included a separate field for micro and small firms (Pieterse et al. 2018: 11–12). Medium to large firms, however, still split this specific variable into the property, plant, and equipment fields and other fixed assets separately. Since 2016, the ITR14 form requires all firms to report vehicles as a separate field. In Table 6, we use an example in which all firms report a value of *Y* in this field. The amount of this field would have been included in the property, plant, and equipment fields of previous versions and, as such, we construct the final property, plant, and equipment measure as in Table 6, as the sum of the property, plant, and equipment fields and the vehicles field.

Table 5: Example of raw ITR14 data

Firm ID (firm category)	Property, plant, and equipment (micro)	Vehicles	Property (medium to large)	Plant (medium to large)	Other fixed assets (medium to large)
1 (micro)	*X1*	—	*NA*	*NA*	*NA*
2 (small)	*X2*	*Y2*	*NA*	*NA*	*NA*
3 (medium to large)	*NA*	*Y3*	*A*	*B*	*C*

Note: this table illustrates the nature of capital variables before harmonization. 'NA' indicates an empty field where the firm was not required to submit this information. '—' indicates a missing field where the firm was required to submit this information but did not or reported a zero entry.

Source: authors' illustration.

[12] Details of these changes are available on request.

Table 6: Example of harmonized ITR14 data

Firm ID (firm category)	Property, plant, and equipment	Other fixed assets
1 (micro)	X1	NA
2 (small)	X2+Y2	NA
3 (large)	A+B+Y3	C

Note: this table illustrates the harmonization of capital variables. 'NA' indicates an empty field where the firm was not required to submit this information.

Source: authors' illustration.

4.1.1.1 Changes in capital stock

Capital stock in the CIT-IRP5 panel is captured by the *k_ppe* variable. The *k_ppe* variable is derived from the balance sheet section on the IT14 and IT14 forms. On the IT14 form, all firms, regardless of their size, completed the same fields. The fields of interest include fixed property amount, fixed assets (plant and equipment) amount, and other fixed assets amount. In the IT14 data, the *k_ppe* variable is created by aggregating the property amount and the fixed assets amount for all firms. The ITR14 form allows smaller firms to provide less detailed information. Where larger firms are required to submit separate information for property and for plant and equipment, smaller firms only need to submit a single field for property, plant, and equipment. Also, smaller firms no longer need to provide information on other fixed assets.

The creation of the *k_ppe* variable across the different versions of the panel has remained the same. However, the values of the *k_ppe* variable are inconsistent across versions. This observed inconsistency appears to be to changes in the form and changes in the underlying data, specifically the inclusion of the vehicles field.

The vehicles field was introduced on the ITR14 form in 2016, and all company types are required to report this amount. This change necessitates the inclusion of the vehicles amount in creating the *k_ppe* variable for all company types. In CIT-IRP5 panel version 3.0, the vehicles amount is not included in the *k_ppe* variable. Therefore, researchers should add the vehicles amount to *k_ppe* when using CIT-IRP5 version 3.0. The *k_ppe* variable in the CIT-IRP5 panel version 4.0, on the other hand, contains the vehicles amount. Adding the vehicles amount to *k_ppe* in CIT-IRP5 version 4.0 will result in double counting. An encoding issue in the raw data further resulted in a portion of capital stock not being counted for medium to large firms in the version 3.0 data; this counting issue is resolved in version 4.

4.1.1.2 Changes in cost of sales

The cost of sales variable is not consistent between versions 2.0 and 3.0 of the panel because of a change in aggregation in the underlying data. On the ITR14 form, only micro, body corporate, and share block companies report the cost of sales amount. Other types of companies report purchase amounts, opening stock, and closing stock. The raw data received from SARS in versions 3.0 and 4.0 contain one cost of sales variable that is inclusive of stock adjustments for larger firms. This field stands in contrast to version 2.0, where these adjustments were not made. It should be noted that this change only holds for ITR14 data after 2012. We therefore include an alternative cost of sales measure (*g_cos2*) that is comparable to the non-stock adjusted cost variables of version 2.0. In general applications, however, the *g_cos* variable is appropriate as the stock adjusted input measure.

4.1.1.3 Constructing balanced books

Researchers should be aware that balanced books cannot be constructed by simply aggregating the fields belonging to different categories, as in Appendix Table A1. First, SARS already construct control totals for specific fields; that is, double counting will necessarily occur if researchers sum all income variables as the control total will be included. Second, we also aggregate information so that specific variables reflect the same economic information across forms and time. As an example, total employment cost is based on a single variable for micro firms, whereas it is based on salaries and wages (including medical, pension, and provident fund contributions) for small businesses. The total employment cost variables for large firms include all fields listed as employee expenses in the ITR14 form. We further construct a variable that is directly comparable to the wages and medical contributions made to small firms by summing wages (*x_ml_wages*), medical payments (*x_ml_medsh*), and pension schemes (*x_ml_pension*). The researcher should therefore carefully go over the construction of variables if they wish to construct balanced books based on the non-control totals in the data.

4.2 Employment data

The employee data come from the IRP5/IT3(a) certificates (henceforth IRP5) submitted to SARS by entities registered for PAYE. Each PAYE entity has a PAYE reference number that is linked to a CIT tax reference number allowing us to merge the two datasets. The current version of the firm-level employment data does not include aggregate information on the employment tax incentives due to data reporting errors (see Ebrahim et al. 2017) or employment by income category due to low usage by researchers and the ease with which researchers can create these variables if needed. The present section is only concerned with the construction of an aggregate firm-level employment measure. The construction of the individual panel is described by Ebrahim and Axelson (2019) and updates to that dataset will be discussed in a separate forthcoming paper by the authors (Ebrahim et al. 2021). We discuss the advantages of specific employment measures in Section 4.2.4.

4.2.1 Identifying employees

4.2.1.1 Employment measures

We identify employees based on types of earnings using four approaches. The 'forms' approach counts all forms submitted as an employee while the 'a3601' approach only counts a form as an employee if non-missing and positive general salary income is reported.[13] It should be noted that this income code was combined with the pension income code for certain years (for details, see Kerr 2016). The forms approach is considered very broad as it counts every form as an employee whereas the a3601 approach is very narrow as it excludes employees with different types of incomes. A form is counted as an employee where at least one of the income codes listed are strictly positive and non-missing.

Employees are also identified with a combination of different income sources. In Table 7, we show the employee income codes that identify an employee under the Kerr approach (see Kerr 2020) and the Pieterse approach (see Pieterse et al. 2018). The Pieterse approach corresponds to the *empl* measure in the previous versions of the panel. The variable was renamed to accurately reflect the fact that it is an employment measure based on a set of assumptions.

[13] This corresponds to the 3601-income source code in the SARS data.

Table 7: Construction of income variables

Measure	Income codes used
forms	Created based on the number of individuals reported at the firm and not dependent on the income codes
a3601	3601
ptrs	3601 3605 3606 3607 3615 3616 3701 3703 3707 3717 3718 3801 3802 3804 3805 3806 3807 3808 3809 3810 3813 3814 3815 3816 3820 3821
kerr	3601 3605 3606 3607 3701 3702 3703 3704 3707 3708 3709 3710 3711 3712 3713 3714 3715 3717 3718 3751 3752 3753 3757 3763 3764 3765 3768 3801 3802 3803 3804 3805 3808 3809 3810 3813 3814 3815 3816 3820 3821 3852 3855 3856 3858 3860 3863 3865

Note: the table includes the income sources codes used in the construction of four approaches to identifying employees.

Source: authors' calculations based on Kerr (2020) and Pieterse et al. (2018).

Each measure has its own benefits and challenges, and we provide guidance to researchers at the end of this subsection.

4.2.1.2 Nature of person

We further aggregate employment information based on the classification of the entity connected to the identification. At present, we only classify the entity as a natural person if the 'nature of person' field on the IRP5 form corresponds to the code that indicates the identifier comes from an identification document or passport number.[14]

We use a basic imputation to categorize a unique identifier to belong to a natural person if that identifier was ever associated with a natural person across the panel. All forms before 2010 are counted as natural persons if they are not observed in the data after 2009, as there is no available nature of person data available before the 2010 data. We also include firm-level employment measures based on a non-imputed nature of person and an ignored nature of person. The non-imputed nature of person uses the nature of person information as given, while the ignored nature of person assigns all IRP5 forms as natural persons. The latter measure will count every single form associated with a firm based on the income measures discussed in the previous section on employment measures. In this context, the ignored nature of person forms measure will count every form in the firm based on the weighting process discussed in Section 4.2.2. This approach will generally include pensioners, and trusts will be included in the measure and should therefore generally be avoided. We list these measures in Table 8 and discuss the preferred employment measures in Section 4.2.4.

Table 8: Nature of person indicators

Nature of person indicator	Description
none	Always count as natural person
basic	Assumes a form belongs to a natural person if ever seen belonging to a natural person; this is the approach used in v2.0 and v3.0.
given	Use SARS identifier as given; treat missing nature of person as missing

Source: authors' classification.

[14] We currently exclude categories B, an individual without an identity or passport number, and C, the director of a private company of member of close corporation. At present, categories B and C are excluded because of noise observed in the data. We also exclude categories D, a Trust, E, a company or close corporation, F, a partnership, G, a corporation, H, a personal service provider, and N, a pensioner.

4.2.2 Weighting employees

We weight employment data in three ways: counting, days worked weight, and periods weight. In all cases, IRP5 forms are aggregated to the CIT tax reference entity and not the PAYE entity. We therefore will not double-count employees who have overlapping periods worked at different PAYE entities belonging to the same firm.

4.2.2.1 Days worked weight

We construct the days worked weighted employment aggregate using the start and end dates reported on the IRP5 form and link it to the firm's financial year. The start and end dates correspond to the first and last dates of the employee's tax period in the year of assessment. SARS allows the start date to be at least 1 January of the previous year and the end date to be at most 30 April of the current year of assessment (SARS 2020c).[15]

We construct the number of days worked in a firm's previous financial year as in Equation (1), where we subtract an individual's start date from the earliest date between the firm's financial year start date and the individual's employment end date. In this way we ensure that workers whose employment date ends before the firm's financial year start date will be included.[16] Similarly, we construct the days worked in the next year by subtracting the last date between the individual's starting period of employment and the firm's financial year end from the individual's employment end date as in Equation (2). To calculate the number of dates worked in the actual firm's financial year corresponding to the IRP5 financial year, we subtract the first date between the firm's financial year end and individual's employment end date from the latest date between the firm's financial year start and the individual's employment start date, as in Equation (3). The number of days worked in the financial year is calculated as in Equation (4), where we count the number of days worked assigned to the previous year based on the next year's IRP5 data, the number of days worked assigned to the current year based on the corresponding year's IRP5 data, and the number of days assigned to the next year based on the previous year's IRP5 data.

$$previous_t = min(Firm\ year\ start_t, Individual\ end\ date_t) - Individual\ start\ date_t \quad \text{if } previous_t \geq 0 \qquad (1)$$

$$next_t = Individual\ end\ date_t - Max(Firm\ year\ end\ date_t, Individual\ start\ date_t) \quad \text{if } next_t \geq 0 \qquad (2)$$

$$in_t = min(Firm\ year\ end_t, End\ date_t) - max(Firm\ start\ date_t, Start\ date_t) + 1 \quad \text{if } in_t \geq 0 \qquad (3)$$

[15] Kerr (2020) discusses valid start and end dates and shows that in the majority of cases where the reported start period is associated with a day before the start of the tax year it occurs in mid-February.

[16] A worker who works for a firm with financial year ending on 31 December 2015 will have tax returns in 2015 submitted for the year ending on 28 February 2015. If this worker only worked from 1 March 2014 to 30 November 2014, they would be included in the 2015 tax returns despite not corresponding with any transactions in the firm's financial year. In the updated version, these workers are completely reallocated to the previous year.

$$days\ worked\ in\ firm_t = previous_{t+1} + in_t + next_{t+1} \quad (4)$$

The *days weight* of each worker is calculated as in Equation (5) by dividing the total number of days worked in the firm by the total number of days in the firm year. *I* is a binary indicator with a value of one where the worker assigned to the form has the relevant employee identifier; *N* is a similar binary indicator reflecting whether the form satisfies the nature of person property. We limit the weight to the [0,1] interval.

$$days\ weight_t(I,N) = min\left(\frac{1}{Days\ in\ firm\ year_t}(Days\ worked\ in\ firm_t) \times I \times N, 1\right) \quad (5)$$

In the previous versions of the panel, a coding error dropped workers who were never present in the period of the firm that overlaps with the tax year. In Section 5.3, we show that this number is likely low as the updated employment measures are not substantially different from those in version 2.0.

4.2.2.2 Periods weight

The *periods weight* calculation is based on the reported 'periods in year of assessment' and 'total periods worked' fields in the IRP5 form. These fields refer to the total number of pay periods worked by the employee in a tax year (SARS 2020c). While the data available for this measure is less precise than the data available for days worked, it is more readily available for earlier periods and used by SARS in the calculation of annual equivalent tax rates (SARS 2020c).

The calculation of periods weighted data is substantially improved from previous versions through the implementation of temporal information. The *periods weight* in previous versions was calculated as in Equation (6), with the aggregation process assuming that the worker works this weight on every day.

$$given\ periods\ weight_t = \frac{Periods\ worked_t}{Total\ periods\ in\ year\ of\ assessment} \quad (6)$$

$$periods\ days_t = Periods\ weight_t \times Days\ in\ tax\ year_t \quad (7)$$

The updated approach constructs a days worked measure based on the *periods weight* as in Equation (7). Where the weight is below a value of one, we set the start and end date of the employee as in Equations (8) and (9), respectively. We then calculate the total number of days worked in the firm as described in the earlier section on employment measures.

$$start\ date_t = \begin{cases} Tax\ year\ start\ date_t & \text{if } Employee\ in\ firm\ in\ t-1\ and periods\ weight_t < 1 \\ Tax\ year\ end\ date_t - Periods\ days_t & \text{if } Employee\ in\ firm\ in\ t+1\ and periods\ weight_t < 1 \end{cases} \quad (8)$$

$$end\ date_t = \begin{cases} Tax\ year\ start\ date_t + Periods\ days_t & \text{if } Employee\ in\ firm\ in\ t-1\ and periods\ weight_t < 1 \\ Tax\ year\ end\ date_t & \text{if } Employee\ in\ firm\ in\ t+1\ and periods\ weight_t < 1 \end{cases} \quad (9)$$

These dates are not defined when the periods weight is below a value of one and the worker is seen both the next and previous tax years. Where workers work with a weight less than a value of one in sequential years, we weight the relative weights of different years by overlap with the financial year of the firm and the IRP5 tax year.

4.2.3 Aggregating employees and measures included

We aggregate employees in each tax paying entity for the corresponding financial year as in Equation (10) where $W \in Periods, Days, Periods\ count, Days\ count$ is the weighting type, $I \in Kerr, Pieterse, a3601, Form$ refers to the employee identifier, and $N \in none, basic, given$ refers to the nature of person calculation. The 'count weights' simply counts each employee as a value of one if they have the available data used to calculate the weight.

$$irp5_t(W, I, N) = \sum_{i=1}^{Forms \in J_t} W_{i,t}(I_{i,t}, N_{i,t}) \quad (10)$$

If all combinations of employment identification, natural person identification, and weighting processes were used, the dataset would contain 36 highly correlated employment measures. All 36 measures are included in the IRP5 firm panel, available to researchers. In the data available to researchers, we limit the selection to the 'days weight', 'periods weight', and 'period weight count weighting' aggregates using the basic nature of person imputation.

4.2.4 Appropriate employment measures

In the present paper, we use the basic nature of person imputation, weighted using the periods data, and using Kerr's (2020) measure of employment income (*irp5_kerr_weight_b*). While the periods data are less granular than the data for days worked, it does appear to be more complete in earlier years allowing for a better comparison across time. The data for days worked may be subject to more error especially if the end of financial years or tax years are around weekends. This should not cause a major problem, but it may result in some employees being given a weight of 0.9945 instead of 1 for certain years. If it is considered simpler to count payment periods than days, it is arguably easier to count payment periods rather than days from the firm's perspective. The question is then whether firms stating, for example, that an employee worked for one payment period reflects a day, month, or year. We do not attempt to answer this question here and

researchers are encouraged to confirm that their results are consistent for the different measures of employment. Many of these measures are included in the final panel but more are available in the IRP5 firm panel should researchers want to merge additional variables.

5 Characteristics of CIT-IRP5 version 4

This section provides descriptive statistics of the CIT-IRP5 panel. Section 5.1 shows the availability of firms by key variables as well as the underlying data match rates in version 4. Section 5.2 compares the aggregates implied by version 4.0 with the aggregate statistics of comparable data by Stats SA, and Section 5.3 compares key variables in version 4.0 with those in versions 2.0 and 3.0.

5.1 Availability of firms by key variables and matched data

Table 9 shows the availability of firms with respect to key variables. We show how the number of tax paying entities in the CIT panel has generally increased from around 750,000 in 2010 to above 900,000 in 2017, with the numbers in 2018 being below 800,000 likely because of lags in submissions. As in previous versions, the majority of firms report missing sales data in 2008. We consider a firm as having valid data for a given variable if it reports a non-missing value that is greater than zero in the field. These restrictions stand in contrast to Pieterse et al. (2018) where availability was separated into non-missing and greater-than-zero categories. This reporting change is due to post-2014 extractions encoding missing data as zeros.

Around 62 per cent of non-dormant firms have valid sales data, with the proportion of non-dormant firms reporting this number as increasing from about 57 per cent in 2010 to around 67 per cent in 2018. The number of firms with valid cost of sales data is substantially lower, ranging from 38 per cent in 2010 to 45 per cent in 2017. Firms with valid capital stock data start from a low of 44 per cent in 2010 and reach around 60 per cent of non-dormant firms in 2018. Conditioning on firms with capital and labour data, however, results in a loss of around 10 per cent of firms compared with the cost of sales data. As in previous versions, the largest drop in firms with valid data comes from employment. Our match rate for firms with labour data and availability of firms with labour data is entirely consistent with the match rates described in Pieterse et al. (2018).

Table 10 shows the number of firms matched from IRP5, customs, and VAT data. As seen, the absolute number of firms matched to the CIT data is increasing in the customs and IRP5 datasets, although the proportion of CIT firms matched is slightly decreasing. The number of firms in the VAT data appears to be falling over time with significant drop-off in 2018. We have no testable hypothesis for this trend at present. The proportion of matches in all cases appears to be like those of previous versions of the data. Tables 9 and 10 indicate that, at least regarding matched data and availability of data, there were no major changes in the data from previous versions. Researchers should thus be careful about the same biases described in Pieterse et al. (2018).

Table 9: Firms by available data

	2008[a]	2009[a]	2010	2011	2012	2013	2014	2015	2016	2017	2018[b]	Total[c]
CIT firms	690,249	595,980	754,750	757,522	813,419	833,077	855,929	873,371	918,050	908,405	788,527	7,310,503
Dormant firms	—	—	183,362	213,230	273,665	306,092	332,551	345,932	382,754	381,209	322,211	2,418,795
Proportion of total firms	—	—	24.3%	28.1%	33.6%	36.7%	38.9%	39.6%	41.7%	42.0%	40.9%	33.1%
Non-dormant firms	690,249	595,980	571,388	544,292	539,754	526,985	523,378	527,439	535,296	527,196	466,316	4,891,708
Proportion of total firms	—	—	75.7%	71.9%	66.4%	63.3%	61.1%	60.4%	58.3%	58.0%	59.1%	66.9%
Sales	7,019	316,761	328,517	326,899	332,802	338,263	343,728	348,816	356,990	355,537	320,446	3,048,313
Proportion of non-dormant firms	[1.0%]	[53.1%]	[57.5%]	[60.1%]	[61.7%]	[64.2%]	[65.7%]	[66.1%]	[66.7%]	[67.4%]	[68.7%]	[62.3%]
Cost of sales	4,302	209,909	218,558	219,037	223,443	225,452	231,066	235,837	243,220	243,735	220,043	2,050,257
Proportion of non-dormant firms	[0.6%]	[35.2%]	[38.3%]	[40.2%]	[41.4%]	[42.8%]	[44.1%]	[44.7%]	[45.4%]	[46.2%]	[47.2%]	[41.9%]
With sales data	4,251	208,785	217,457	218,063	222,400	224,292	229,824	234,612	241,866	242,407	218,876	2,039,706
Proportion of restricted firms	|98.8%|	|65.9%|	|66.2%|	|66.7%|	|66.8%|	|66.3%|	|66.9%|	|67.3%|	|67.8%|	|68.2%|	|68.3%|	|66.9%|
Proportion of non-dormant firms	[0.6%]	[35.0%]	[38.1%]	[40.1%]	[41.2%]	[42.6%]	[43.9%]	[44.5%]	[45.2%]	[46.0%]	[46.9%]	[41.7%]
Capital stock	7,288	335,539	339,706	332,876	333,024	329,176	330,257	330,673	331,910	324,418	286,622	2,987,579
Proportion of non-dormant firms	[1.1%]	[56.3%]	[59.5%]	[61.2%]	[61.7%]	[62.5%]	[63.1%]	[62.7%]	[62.0%]	[61.5%]	[61.5%]	[61.1%]
With sales and cost of sales data	3,449	178,320	184,320	183,048	185,268	186,445	189,961	192,773	196,302	194,679	173,380	1,691,116
Proportion of restricted firms	|81.1%|	|85.4%|	|84.8%|	|83.9%|	|83.3%|	|83.1%|	|82.7%|	|82.2%|	|81.2%|	|80.3%|	|79.2%|	|82.9%|
Proportion of non-dormant firms	[0.5%]	[29.9%]	[32.3%]	[33.6%]	[34.3%]	[35.4%]	[36.3%]	[36.5%]	[36.7%]	[36.9%]	[37.2%]	[34.6%]
Labour	156,829	150,389	162,345	165,626	168,921	169,958	170,993	176,403	180,188	181,233	—	1,526,056
Proportion of non-dormant firms	[22.7%]	[25.2%]	[28.4%]	[30.4%]	[31.3%]	[32.3%]	[32.7%]	[33.4%]	[33.7%]	[34.4%]	—	[31.2%]
With sales, cost of sales, and capital data	1,475	103,893	109,083	111,586	113,587	114,588	116,811	120,267	122,467	122,830	—	1,035,112
Proportion of restricted firms	|42.8%|	|58.3%|	|59.2%|	|61.0%|	|61.3%|	|61.5%|	|61.5%|	|62.4%|	|62.4%|	|63.1%|	—	|61.2%|
Proportion of non-dormant firms	[0.2%]	[17.4%]	[19.1%]	[20.5%]	[21.0%]	[21.7%]	[22.3%]	[22.8%]	[22.9%]	[23.3%]	—	[21.2%]

Note: this table shows the number of CIT firms by availability of data for 2008–2018. ‘CIT firms’ represents the total number of unique tax reference numbers for the given year. ‘Dormant firms’ represents the total number of firms satisfying any dormancy criteria in a given year. ‘Non-dormant firms’ represents the total number of firms not satisfying the dormancy criteria in Section 3.7. Sales, cost of sales, capital stock, and labour categories reflect the total number of non-dormant firms with greater-than-zero and non-missing data for the relevant variable, respectively. Sales, Cost of sales, and Capital stock are measured by the g_sales, g_cos, and k_ppe fields, respectively. Labour is measured by the Kerr weighted employment measure using the basic nature of person imputation as described in Section 4.2. ‘Proportion of total firms’ represents the proportion of firms satisfying the dormancy criteria of the row as a percentage of all CIT firms denoted with no brackets. ‘Proportion of non-dormant firms’ represents the proportion of firms with the available field reported in the first column as a percentage of all non-dormant firms denoted with square brackets [ ]. ‘Proportion of restricted firms’ is the percentage of firms with the available data in the field preceding it; for example, the proportion of restricted firms in the non-zero labour section is the proportion of non-dormant firms with sales, cost of sales, capital, and labour data as a proportion of non-dormant firms with sales, cost of sales, and capital data denoted with straight brackets | |. [a] The dormancy indicators are

not reported for 2008 and 2009 due to differences in availability in the underlying data. [b] At the time of writing, an extraction concern with the 2018 and 2019 labour data necessitated its exclusion from the panel. [c] The total field excludes 2008 because of the general unavailability of data in the year and excludes 2018 due to unavailability of labour data. '—' indicates unreported statistics due to the reasons specified in notes a, b, and c.

Source: authors' calculations based on CIT-IRP5 firm-level panel data v4.0 (National Treasury and UNU-WIDER 2021a).

Table 10: Firms by matched data

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
CIT firms	690,249	595,980	754,750	757,522	813,419	833,077	855,929	873,371	918,050	908,405	788,527
IRP5 firms	226,089	226,675	237,158	239,905	244,498	247,756	252,992	259,887	266,409	274,350	289,103
Matched to CIT	166,617	159,286	175,871	178,218	181,881	183,252	184,068	189,414	193,605	195,049	187,157
Proportion of CIT firms	24.1%	26.7%	23.3%	23.5%	22.4%	22.0%	21.5%	21.7%	21.1%	21.5%	23.7%
Proportion of IRP5 firms	73.7%	70.3%	74.2%	74.3%	74.4%	74.0%	72.8%	72.9%	72.7%	71.1%	64.7%
Customs firms	NA	29,220	51,058	53,807	54,445	58,228	59,229	59,671	60,549	60,976	60,914
Matched to CIT	NA	20,670	34,950	36,433	37,841	39,459	39,721	39,898	39,823	38,746	33,333
Proportion of CIT firms	NA	3.5%	4.6%	4.8%	4.7%	4.7%	4.6%	4.6%	4.3%	4.3%	4.2%
Proportion of customs firms	NA	70.7%	68.5%	67.7%	69.5%	67.8%	67.1%	66.9%	65.8%	63.5%	54.7%
VAT firms	659,144	620,636	575,373	517,136	504,632	493,086	488,529	485,738	490,037	498,373	249,431
Matched VAT firms	343,400	313,930	337,975	317,651	314,914	309,272	307,242	308,758	304,344	303,572	249,431
Proportion of CIT firms	49.8%	52.7%	44.8%	41.9%	38.7%	37.1%	35.9%	35.4%	33.2%	33.4%	31.6%
Proportion of VAT firms	52.1%	50.6%	58.7%	61.4%	62.4%	62.7%	62.9%	63.6%	62.1%	60.9%	100.0%

Note: 'NA' indicates no customs data are available for 2008.

Source: authors' calculations based on CIT-IRP5 firm-level panel data v4.0 (National Treasury and UNU-WIDER 2021a).

5.2 Aggregate data

In Table 11, we compare total sales in the CIT data with total sales reported in the quarterly financial statistics (QFS) of Stats SA (2020). The former refers to the sales amount reported in the income statement of the firm in the IT14 and ITR14 forms; this information is not necessarily the same as the gross income amount variable that includes sales as well as amounts from other income sources (SARS 2020b).

Table 11 uses the Budlender and Ebrahim (2020) industry classification, meaning it is not directly comparable to the decomposition reported in Pieterse et al. (2018). The updated industry classification variable allows for an increase of 20 per cent of output to assigned industries compared with Pieterse et al. (2018) for 2009–2013 whereas the updated panel's output aggregate hovers around 100 per cent compared with previous versions. In this context, Budlender and Ebrahim's (2020) industry classification does appear to classify more firms more consistently over time.

The sample restrictions in Table 11 are not directly comparable to Pieterse et al. (2018) as later vintages of the underlying data treat missing observations as zero, meaning that the aggregates in the table require all fields to be non-missing and non-zero when restricting the sample. The restriction used in Table 11 is more limiting than the Pieterse et al. (2018) restrictions that allowed firms with values of zero for specific fields to be counted. Our sales aggregate for all industries while limiting the data gives aggregates around 90 per cent of that of version 2. Note that using the same restriction as Pieterse et al. (2018) yields an aggregate 2 per cent higher in the current data; this is consistent with the hypothesis that reporting lags are behind the general drop seen in the aggregates of years close to the end of the panel. The total sales of the sample with all key variables constitute around 80 per cent of the QFS sales aggregates. This aggregate excludes the agricultural sector as it is not available in the QFS. In Pieterse et al. (2018) this average was around 83 per cent but ranging from 80 to 85 per cent. Where we include firms reporting zero values for sales, cost of sales, capital stock, and employment, as having non-missing data representation would increase to around 100 per cent of the QFS aggregate. The figures with less severe restrictions are available on request. The dramatic shifts in QFS representation seen in the Finance sector and the Community and Social Services sector are consistent with those found by Pieterse et al. (2018).

Table 11: Total sales by industry in the QFS and CIT-IRP5 panel v4.0, Rand millions

Industry	Data	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Agriculture	No restrictions	146,958	184,608	184,608	187,344	230,328	257,794	323,136	339,430	418,533	398,033
	Has all key variables	91,087	129,311	129,188	132,625	165,911	190,936	223,977	256,181	275,548	316,348
	Percentage of no restrictions	62.0%	70.0%	70.0%	70.8%	72.0%	74.1%	69.3%	75.5%	65.8%	79.5%
Mining	No restrictions	273,344	375,485	375,485	336,920	500,533	583,174	599,993	595,996	616,858	657,025
	Percentage of QFS	NA	[118.7%]	[118.7%]	[82.5%]	[108.3%]	[125.5%]	[113.4%]	[114.8%]	[117.4%]	[109.5%]
	Has all key variables	72,446	274,028	273,704	235,744	302,743	361,672	415,670	401,584	415,027	439,231
	Percentage of no restrictions	26.5%	73.0%	72.9%	70.0%	60.5%	62.0%	69.3%	67.4%	67.3%	66.9%
	Percentage of QFS	NA	[86.6%]	[86.5%]	[57.7%]	[65.5%]	[77.8%]	[78.6%]	[77.3%]	[79.0%]	[73.2%]
	QFS	NA	316,285	316,285	408,239	462,196	464,726	529,124	519,185	525,272	600,020
Manufacturing	No restrictions	1,736,800	1,942,089	1,942,089	2,151,063	2,524,453	2,884,890	3,082,693	3,100,080	3,189,007	3,355,047
	Percentage of QFS	NA	[128.7%]	[128.7%]	[129.1%]	[133.1%]	[143.0%]	[135.5%]	[134.9%]	[136.8%]	[134.1%]
	Has all key variables	1,502,577	1,532,484	1,531,819	1,715,770	2,037,022	2,323,978	2,529,172	2,566,236	2,693,316	2,810,900
	Percentage of no restrictions	86.5%	78.9%	78.9%	79.8%	80.7%	80.6%	82.0%	82.8%	84.5%	83.8%
	Percentage of QFS	NA	[101.5%]	[101.5%]	[103.0%]	[107.4%]	[115.2%]	[111.2%]	[111.7%]	[115.6%]	[112.4%]
	QFS	NA	1,509,477	1,509,477	1,665,577	1,896,778	2,017,591	2,274,545	2,297,348	2,330,360	2,501,004
Electricity, gas, and water	No restrictions	20,315	47,080	47,080	139,891	176,315	182,579	206,883	278,664	266,682	276,464
	Percentage of QFS	NA	[53.0%]	[53.0%]	[123.6%]	[131.7%]	[122.8%]	[121.7%]	[150.7%]	[126.5%]	[115.9%]
	Has all key variables	14,416	17,241	17,229	118,694	155,264	153,480	171,064	182,913	221,672	218,933
	Percentage of no restrictions	71.0%	36.6%	36.6%	84.8%	88.1%	84.1%	82.7%	65.6%	83.1%	79.2%
	Percentage of QFS	NA	[19.4%]	[19.4%]	[104.9%]	[116.0%]	[103.2%]	[100.6%]	[98.9%]	[105.1%]	[91.8%]
	QFS	NA	88,893	88,893	113,166	133,842	148,686	170,048	184,856	210,888	238,529
Construction	No restrictions	261,437	585,924	585,924	261,299	296,846	352,080	415,758	440,273	477,428	484,264
	Percentage of QFS	NA	[230.7%]	[230.7%]	[105.1%]	[117.8%]	[133.7%]	[137.5%]	[123.7%]	[119.4%]	[115.2%]
	Has all key variables	179,519	191,349	190,760	193,165	218,922	261,175	302,881	327,733	357,697	363,735
	Percentage of no restrictions	68.7%	32.7%	32.6%	73.9%	73.7%	74.2%	72.9%	74.4%	74.9%	75.1%
	Percentage of QFS	NA	[75.3%]	[75.1%]	[77.7%]	[86.9%]	[99.2%]	[100.1%]	[92.1%]	[89.4%]	[86.5%]
	QFS	NA	253,958	253,958	248,652	251,888	263,331	302,462	355,977	399,905	420,312

Wholesale, retail, catering, and accommodation	No restrictions	1,039,421	1,275,685	1,275,685	1,461,634	1,663,385	1,866,020	2,023,682	2,125,241	2,293,533	2,433,406
	Percentage of QFS	NA	[79.0%]	[79.0%]	[80.7%]	[79.2%]	[83.5%]	[82.3%]	[79.2%]	[81.4%]	[77.1%]
	Has all key variables	732,807	924,888	923,741	1,115,335	1,250,114	1,472,549	1,584,955	1,719,983	1,888,414	2,057,990
	Percentage of no restrictions	70.5%	72.5%	72.4%	76.3%	75.2%	78.9%	78.3%	80.9%	82.3%	84.6%
	Percentage of QFS	NA	[57.3%]	[57.2%]	[61.6%]	[59.5%]	[65.9%]	[64.4%]	[64.1%]	[67.0%]	[65.2%]
	QFS	NA	1,614,508	1,614,508	1,811,128	2,100,242	2,233,420	2,459,414	2,683,340	2,819,281	3,158,051
Transport, storage, and communication	No restrictions	268,884	532,773	532,773	582,641	632,262	650,092	655,844	709,962	714,153	760,341
	Percentage of QFS	NA	[124.5%]	[124.5%]	[125.4%]	[118.5%]	[111.9%]	[101.2%]	[104.7%]	[100.3%]	[96.0%]
	Has all key variables	129,081	293,468	293,415	347,121	352,574	327,451	334,079	364,148	395,573	432,882
	Percentage of no restrictions	48.0%	55.1%	55.1%	59.6%	55.8%	50.4%	50.9%	51.3%	55.4%	56.9%
	Percentage of QFS	NA	[68.6%]	[68.6%]	[74.7%]	[66.1%]	[56.3%]	[51.6%]	[53.7%]	[55.6%]	[54.7%]
	QFS	NA	427,865	427,865	464,587	533,515	581,170	647,927	678,135	711,913	791,739
Financing, insurance, real estate, and business services	No restrictions	501,515	941,628	941,628	889,586	894,424	694,015	815,933	862,339	918,001	920,477
	Percentage of QFS	NA	[197.1%]	[197.1%]	[190.0%]	[168.9%]	[114.2%]	[107.8%]	[94.8%]	[93.7%]	[88.2%]
	Has all key variables	211,676	381,492	380,912	374,612	342,887	298,376	359,466	389,065	421,053	428,717
	Percentage of no restrictions	42.2%	40.5%	40.5%	42.1%	38.3%	43.0%	44.1%	45.1%	45.9%	46.6%
	Percentage of QFS	NA	[79.8%]	[79.7%]	[80.0%]	[64.8%]	[49.1%]	[47.5%]	[42.8%]	[43.0%]	[41.1%]
	QFS	NA	477,775	477,775	468,311	529,456	607,975	757,179	909,860	979,418	1,043,526
Community and social services	No restrictions	190,445	242,761	242,761	440,186	331,433	370,117	404,772	454,218	508,359	552,934
	Percentage of QFS	NA	[206.3%]	[206.3%]	[298.3%]	[200.5%]	[210.2%]	[223.1%]	[223.8%]	[260.4%]	[251.0%]
	Has all key variables	91,465	113,603	113,289	146,943	189,417	213,226	247,504	279,109	332,120	356,993
	Percentage of no restrictions	48.0%	46.8%	46.7%	33.4%	57.2%	57.6%	61.1%	61.4%	65.3%	64.6%
	Percentage of QFS	NA	[96.5%]	[96.3%]	[99.6%]	[114.6%]	[121.1%]	[136.4%]	[137.5%]	[170.1%]	[162.0%]
	QFS	NA	117,674	117,674	147,583	165,277	176,072	181,448	202,940	195,246	220,303
Economy	No restrictions	4,442,199	6,128,061	6,128,061	6,450,590	7,249,986	7,840,761	8,528,698	8,906,207	9,402,559	9,837,994
	Assigned industries	4,439,119	6,128,033	6,128,033	6,450,564	7,249,979	7,840,761	8,528,694	8,906,203	9,402,554	9,837,991
	QFS industries	4,292,161	5,943,425	5,943,425	6,263,220	7,019,651	7,582,967	8,205,558	8,566,773	8,984,021	9,439,958
	Percentage of QFS	NA	[123.67%]	[123.67%]	[117.57%]	[115.58%]	[116.79%]	[112.06%]	[109.39%]	[109.93%]	[105.20%]
	Has all key variables	3,033,031	3,857,882	3,857,882	4,383,911	5,018,798	5,611,203	6,172,186	6,491,948	7,005,004	7,431,545
	Percentage of no restrictions	68.28%	62.9%	62.9%	67.96%	69.22%	71.56%	72.37%	72.89%	74.50%	75.54%

Assigned industries	3,032,260	3,857,864	3,857,864	4,383,895	5,018,798	5,611,203	6,172,182	6,491,944	7,004,999	7,431,545
Percentage of no restrictions	68.31%	62.9%	62.9%	67.96%	69.23%	71.56%	72.37%	72.89%	74.50%	75.54%
QFS industries	2,940,827	3,728,553	3,728,553	4,251,133	4,852,776	5,420,097	5,948,085	6,235,522	6,729,361	7,115,023
Percentage of no restrictions	68.52%	62.7%	62.7%	67.87%	69.13%	71.48%	72.49%	72.79%	74.90%	75.37%
Percentage of QFS aggregate	NA	[77.6%]	[77.6%]	[79.8%]	[79.9%]	[83.48%]	[81.23%]	[79.62%]	[82.34%]	[79.29%]
QFS aggregate	NA	4,806,435	4,806,435	5,327,243	6,073,194	6,492,971	7,322,147	7,831,641	8,172,283	8,973,484

Note: QFS, quarterly financial statistics. Percentages in brackets [ ] are sales aggregates as a proportion of QFS aggregates; percentages without brackets are sales aggregates compared with unrestricted aggregates. 'NA' indicates that no data, or in the case of the QFS data no comparable data, exist. '—' indicates that data do exist, but due to concerns with the labour data the aggregates are not included. Few firms exist with an unassigned industry not reported in this table due to data security.

Source: authors' calculations based on CIT-IRP5 firm-level panel data v4.0 (National Treasury and UNU-WIDER 2021a) and the QFS (Stats SA 2020).

5.3 Comparing CIT-IRP5 version 4.0 with previous versions

The values of variables in version 4.0 of the dataset may not always be the same as in previous panels. The differences may arise either from improvements in the cleaning process or from changes to the underlying data. In some cases, differences may also be explained by revisions made to previously submitted information. Table 12 presents the differences in key variables between version 4.0 and the previous versions. We construct the mean as the difference of each key variable in firm i for year t in version 4.0 and $z \in 2,3,$ as in Equation (11).

$$X = E\left[X_{i,t}^{v.4} - X_{i,t}^{v.z}\right] \tag{11}$$

As expected, there are statistically significant differences in the employment variable (*irp5_weight_b*) because of improvements to the cleaning process of employment indicators in the dataset. Note that this difference is extremely small in general, with employment figures being on average between −1 and +1 of the version 2.0 values with majority of increases occurring for the years 2009 and 2010, where version 2.0 had issues with employment figures. The measures deviate more from the version 3.0 figures, but generally stay within a range of −3 and +3 on average.

The sales variable (*g_sales*) in version 4.0 is generally not statistically significantly different from the previous versions, except for the tax years 2013 and 2014 (2014 and 2015) compared with version 3.0 (2.0). The changes in the data do not follow a clear within-firm trend, indicating that these differences are likely because of changes in information submitted by firms. This hypothesis is further based on the lower deviation, in absolute terms, in version 3.0 data compared with version 2.0 data. The former is based on information closer to version 4.0 data.

The capital variable (*k_ppe*) is significantly different from version 3.0 across most tax years and only different for the tax years 2009 and 2012 compared with version 2.0. The difference in the capital variable results from changes to the raw data, as received from SARS. The raw extractions for versions 3.0 and 4.0 came with a variable called 'old, fixed assets', a field not available on the forms, which must be part of the *k_ppe* aggregation to make the capital variable more consistent with version 2.0. When creating version 3.0 of the CITIRP5 panel, *k_ppe* did not include the 'old, fixed assets' amount, hence the observed difference in *k_ppe* between versions 4.0 and 3.0. Version 4.0 consists of a variable called *k_ppe_undj*, an aggregation of property amount, plant and equipment amount, and vehicles amount. We adjust the *k_ppe_unadj* by adding to it the 'old, fixed assets' amount to get *k_ppe*. Given that the capital variable in version 3.0 does not include the vehicles amount, we observe a significant difference between *k_ppe_unadj* and *k_ppe* in version 3.0, especially in 2016 and 2017.

Table 12: Differences in key variables across versions

	Version	*g_sales*	*g_cos*	*g_cos2*	Value added	*k_ppe*	*k_ppe_unadj*	Employment
2008	v2	0.073	−90,113*	NA	18,513	−0.0058	−0.0058	14***
		(7,023)	(4,276)	NA	(7,059)	(7,249)	(7,249)	(142,072)
		[6.1]	[3,045,659]	NA	[3,021,274]	[0.43]	[0.43]	[363]
	v3	NA	NA	NA	NA	NA	NA	0.74**
		NA	NA	NA	NA	NA	NA	(157,073)
		NA	NA	NA	NA	NA	NA	[136]
2009	v2	2,120	−7,685	NA	9,913*	31,568***	31,568***	6***
		(312,907)	(207,371)	NA	(230,586)	(262,990)	(262,990)	(138,265)
		[5,696,464]	[5,546,722]	NA	[2,889,397]	[5,647,416]	[5,647,416]	[244]
	v3	−209	−8,930	NA	15,462,846***	6,288*	6,288*	−1.2**
		(71,523)	(201,973)	NA	(205,746)	(253,811)	(253,811)	(148,039)
		[2,211,201]	[4,561,067]	NA	[1298552537]	[1,773,277]	[1,773,277]	[186]
2010	v2	−889,843	6,856	9,098	−1,149,913	−11,081	−2,914,833**	1.1
		(322,826)	(214,868)	(214,896)	(252,308)	(269,921)	(270,062)	(157,520)
		[508,492,114]	[3,844,270]	[39,136,113]	[575,168,846]	[5,585,596]	[688,503,939]	[707]
	v3	0	−2,780	0.0099	1,221,441	2,845,948**	−0.0011*	−2.6*
		(329,549)	(219,163)	(223,275)	(257,744)	(276,061)	(276,273)	(161,730)
		[0]	[38,586,435]	[3.1]	[612,082,719]	[680,965,264]	[0.3]	[539]
2011	v2	3,357	−21,339**	278,017	−11,988	−12,123	−2,295,248***	0.65
		(317,257)	(212,552)	(212,585)	(306,013)	(265,204)	(265,343)	(161,689)
		[3,596,992]	[4,337,997]	[688,929,671]	[12,208,162]	[6,573,892]	[340,775,969]	[665]
	v3	0	−290,517	−0.0082	209,992	2,206,368***	0.0027	−3.4**
		(327,451)	(219,374)	(222,458)	(314,532)	(274,588)	(274,799)	(165,777)
		[0]	[678,186,823]	[4.4]	[566,383,652]	[334,928,116]	[2]	[566]
2012	v2	2,148	−151,300***	−58,540	−108,023	927**	−362,607***	0.12
		(307,831)	(206,837)	(206,937)	(308,164)	(266,458)	(266,561)	(163,059)
		[1,205,996]	[12,187,590]	[22,292,170]	[71,398,751]	[212,342]	[39,292,711]	[581]
	v3	29	−91,724*	−75	487,190	335,503***	−81	−3.5***
		(331,053)	(222,384)	(224,366)	(330,439)	(288,776)	(288,954)	(169,168)
		[40,065]	[24,199,817]	[59,459]	[243,203,898]	[37,750,767]	[54,594]	[535]
2013	v2	13,834**	−185,072***	11,449*	−66,748	−800	−998	−1.1***

		(295,790)	(195,776)	(196,070)	(296,262)	(282,773)	(282,775)	(168,344)
		[3,788,184]	[29,952,664]	[2,795,818]	[60,742,671]	[730,886]	[731,251]	[123]
	v3	823	−191,254***	139	128,341***	239***	64*	−2.1***
		(335,874)	(223,955)	(224,312)	(336,573)	(320,736)	(320,739)	(169,367)
		[394,554]	[28,429,170]	[58,704]	[23,192,310]	[33,251]	[19,884]	[148]
2014	v2	9,801***	76,292	3,189	−82,489	2,172	2,130	–0.19
		(240,614)	(159,239)	(159,475)	(241,007)	(230,596)	(230,596)	(167,731)
		[1,063,380]	[61,232,099]	[865,541]	[49,907,659]	[1,057,425]	[1,057,346]	[221]
	v3	1,147**	−79,163	98	55,110	−5.2	−40	−1.1**
		(338,153)	(227,423)	(227,781)	(338,912)	(319,817)	(319,817)	(169,200)
		[267,974]	[55,934,690]	[115,035]	[45,820,752]	[53,072]	[51,923]	[224]
2015	v2	NA	NA	NA	NA	NA	NA	0.37
		NA	NA	NA	NA	NA	NA	(156,605)
		NA	NA	NA	NA	NA	NA	[234]
	v3	1,532**	−156,144**	−47	108,138**	−75	−98	–0.4
		(339,089)	(229,425)	(229,794)	(339,753)	(316,418)	(316,418)	(171,541)
		[346,395]	[33,684,781]	[251,357]	[27,683,152]	[107,804]	[107,559]	[229]
2016	v2	NA	NA	NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA	NA	NA
	v3	−54,928	−111,811	−6,601	23,181	409,266**	409,257**	–0.089
		(339,156)	(231,156)	(231,519)	(339,881)	(301,174)	(301,174)	(170,154)
		[33,574,316]	[49,781,177]	[3,559,017]	[51,143,809]	[88,226,457]	[88,226,465]	[244]
2017	v2	NA	NA	NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA	NA	NA
	v3	−1,482	−79,770	2,325	53,703	573,200***	573,195***	1.9***
		(315,015)	(215,505)	(215,849)	(315,626)	(271,271)	(271,271)	(159,233)
		[1,847,116]	[37,482,580]	[778,327]	[31,026,690]	[94,959,975]	[94,959,999]	[209]

Note: the number of observations is reported in parentheses ( ) and the standard errors are in brackets [ ]. ***Significant at 1 per cent, **significant at 5 per cent, *significant at 10 per cent. ‘NA’ indicates that no comparable field exists between versions of the dataset.

Source: authors’ calculations based on CIT-IRP5 firm-level panel data v2.0, v3.0, and v4.0 (National Treasury and UNU-WIDER 2019, 2020, 2021a).

In Appendix B we show the log ratio of version 4.0 and previous versions' sales, property, plant and equipment, and employment numbers. In general, the distributions are extremely tightly dispersed around zero, indicating that majority of the differences are due to a few firms instead of general differences. In Appendix Figure B1, the dispersion in the log difference of sales against version 2.0 is highest for 2011–2013, with the generally jagged nature of the dispersion again highlighting that discrete differences exist for specific firms. There is substantial dispersion comparing sales with version 3.0, with a few firms reporting double or half of the amount that they did in version 3.0.[17] There is a peak in 2015 and 2013 around 0.4 that corresponds to a substantial difference of around 50 per cent from the original reported amount for both versions 2.0 and 3.0. This is likely because of reporting lags and other issues related to submitted forms in the year before the final year of the panel, as most firms have submitted but not yet revised their forms.

In Appendix Figure B2, we show similar tight dispersion for capital stock with only 2013 being substantially more widely dispersed for version 2.0. We again see a substantial peak around 50 per cent higher than the capital stock reported for this year. The differences in version 3.0 data follow a similar trend, with extremely jagged peaks at specific intervals. This may be due to revisions.

Appendix Figure B3 shows substantially similar employment numbers for all firms. Compared with version 2.0 figures, we show that where differences exist persistently, they are slightly higher in version 4.0 owing to the updated counting approach that captures more employees in firms not ending in February.

Taken together, the updated panel closely matches the previous data with within-firm differences being minimal or being explained by changes in the underlying data. The relatively clear upward shift for data in the pre-final year of a version—2013 for version 1.0 and 2015 for version 2.0—highlights the importance of changes in submissions across vintages. That is, there will likely be some measurement error in key variables for the year before the end of the final year in question. Specifically, it appears likely that firms will underreport income in the pre-final year of the data. This measurement concern is completely different from the sample concern of late submissions in the final year of the data, where larger firms are less likely to have already submitted returns. In this context, using the updated data to confirm previous results is crucial not only from a quality assurance perspective but also to correct for measurement error.

6 Conclusion

The CIT-IRP5 panel has been used several times for academic and policy research since it was created, which is a testament to the data as a unique and significant source of information for the study of firm behaviour in post-apartheid South Africa. A big part of the improvements in the panel in each version is due to the researchers who have encountered data challenges and proposed solutions or fixes for this version and previous versions of the data.

In this paper, we have presented the updated version of the CIT-IRP5 panel. We discussed substantial improvements incorporated based on the works of Budlender and Ebrahim (2020), Kerr (2020), and Kilumelume et al. (2021), and showed that, despite changes in the underlying

[17] Recall that $\ln(2) \approx 0.693$ and $\ln(1.5) \approx 0.405$.

data, changes in vintages, and several form changes, the updated data largely conform to previous versions and external sources.

Although we do improve on many aspects of the data, many improvements are currently being considered for the next version of the panel. These include addressing some smaller variable inconsistencies, temporal fixes, and improvement to aspects of the trade and VAT data. Researchers are encouraged to continue providing feedback to improve the panel for all users.

References

Budlender, J., and A. Ebrahim (2020). 'Industry Classification in the South African Tax Microdata'. WIDER Working Paper 2020/99. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/856-6

Ebrahim, A., M. Leibbrandt, and V. Ranchhod (2017). 'The Effects of the Employment Tax Incentive on South African Employment'. WIDER Working Paper 2017/5. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2017/229-8

Ebrahim, A., and C. Axelson (2019). 'The Creation of an Individual Panel Using Administrative Tax Microdata in South Africa'. WIDER Working Paper 2019/27. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/661-6

Ebrahim, A., C. Axelson, D. Brink, and G. Bridgeman (2021). 'The Guide to the Individual Panel Version 4.0'. WIDER Technical Note 2021 [forthcoming]. Helsinki: UNU-WIDER.

Kerr, A. (2016). 'Job Flows, Worker Flows, and Churning in South Africa'. WIDER Working Paper 2016/37. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2016/080-5

Kerr, A. (2020). 'Earnings in the South African Revenue Service IRP5 Data'. WIDER Working Paper 2020/62. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/819-1

Kilumelume, M., H. Reynolds, and A. Ebrahim (2021). 'Identifying Foreign Firms and South African Multination Enterprises.' WIDER Technical Note 2021/1. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-1

National Treasury (2018). '2018 Budget Speech 2018'. Delivered by Malusi Gigaba, Minister of Finance, 21 February. Pretoria: National Treasury South Africa.

National Treasury and UNU-WIDER (2019). 'CIT-IRP5 Firm Level Panel 2008–2016 [dataset]. Version 2.0'. Pretoria: South African Revenue Service [producer of the original data], 2018. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2019.

National Treasury and UNU-WIDER (2020). 'CIT-IRP5 Firm Level Panel 2008–2017 [dataset]. Version 3.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2021a). 'CIT-IRP5 Firm Level Panel 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2020. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2021.

National Treasury and UNU-WIDER (2021b). 'Customs Firm Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2020. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2021.

National Treasury and UNU-WIDER (2021c). 'Customs Transaction Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2020. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2021.

Pieterse, D., E. Gavin, and F.C. Kreuser (2018). 'Introduction to the South African Revenue Service and National Treasury Firm-Level Panel'. Firm Level Analysis Using Administrative Record Data, *South African Journal of Economics*, 86(S1): 6–39.

SARS (2019). *Legal Counsel: Value-Added Tax* [online]. South African Revenue Service (SARS). Available at: https://www.sars.gov.za/wp-content/uploads/Ops/Guides/LAPD-VAT-G02-VAT-404-Guide-for-Vendors.pdf (accessed 27 September 2021).

SARS (2020a). *External Guide: How to Complete the Company Income Tax Return eFiling*. Revision 8 [Online]. South African Revenue Service (SARS). Available: https://www.sars.gov.za/wp-content/uploads/Ops/Guides/IT-ELEC-03-G01-How-to-complete-the-company-Income-Tax-return-ITR14-eFiling-External-Guide.pdf (accessed 12 July 2021).

SARS (2020b). *External Guide: How to Complete the Company Income Tax Returns*. Revision 12 [Online]. South African Revenue Service (SARS). Available at: https://www.sars.gov.za/wp-content/uploads/Ops/Guides/IT-GEN-04-G01-How-to-complete-the-Income-Tax-Return-ITR14-for-Companies-External-Guide.pdf (accessed 12 July 2021).

SARS (2020c). *A Step-by-Step Guide to the Employer Reconciliation Process*. Revision 11 [Online]. South African Revenue Service (SARS). Available at: https://www.sars.gov.za/wp-content/uploads/Ops/Guides/EMP-GEN-02-G01-A-Step-by-Step-Guide-to-the-Employer-Reconciliation-Process-External-Guide.pdf (accessed 12 July 2021).

SARS (2020d). *Tax Guide for Small Businesses* [Online]. South African Revenue Service (SARS). Available at: https://www.sars.gov.za/wp-content/uploads/Ops/Guides/LAPD-IT-G10-Tax-Guide-for-Small-Businesses.pdf (accessed 17 July 2021).

SARS (2020e). *Pay As You Earn* [Online]. South African Revenue Service (SARS). Available at: https://www.sars.gov.za/types-of-tax/pay-as-you-earn/ (accessed 12 July 2021).

SARB (2020a). *National Accounts December 2020* [Dataset]. South African Reserve Bank (SARB) Publication. Available at: https://www.resbank.co.za/content/dam/sarb/publications/quarterly-bulletins/download-information-from-xlsx-data-files/2020/dec-2020/National%20Accounts%20December%202020.zip (accessed 17 October 2020).

SARB (2020b). *Business Cycle December 2020*. South African Reserve Bank (SARB) Publication. Available at: https://www.resbank.co.za/content/dam/sarb/publications/quarterly-bulletins/download-information-from-xlsx-data-files/2020/dec-2020/National%20Accounts%20December%202020.zip (accessed 17 October 2020).

Stats SA (2012). *Standard Industrial Classification of all Economic Activities (SIC) Seventh Edition*. Pretoria: Statistics South Africa (Stats SA).

Stats SA (2020). Quarterly Financial Statistics 2008–2020 [datasets]. Pretoria: Statistics South Africa (Stats SA) [Producer of the original data]. Available at: http://www.statssa.gov.za/?page_id=1854&PPN=P0044 (accessed 17 October 2020).

South Africa (1962). Income Tax Act 58 of 1962. Available at: https://www.gov.za/documents/income-tax-act-29-may-1962-0000 (accessed 28 June 2021).

World Bank (2020). World Bank list of economies (June 2020). Available at: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (accessed 27 September 2021).

Appendix A

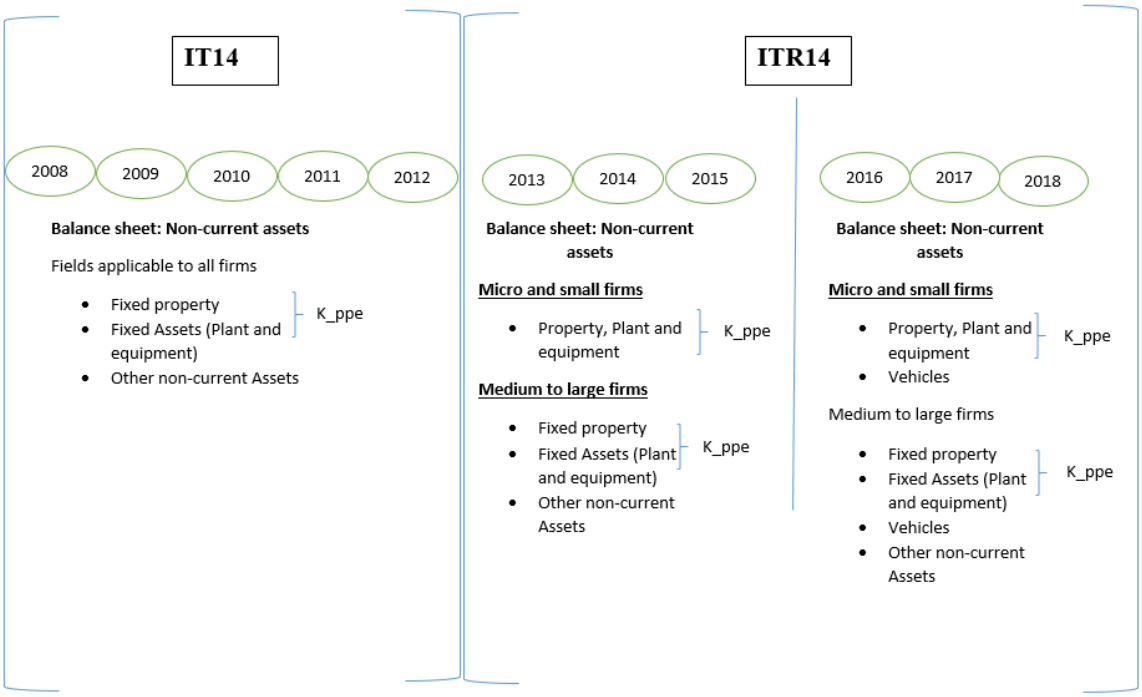
Table A1: CIT-IRP5 firm-level panel variable prefixes

Prefix	Description	Source
c_	Characteristic variable	Based on yes/no, categorical, or information not provided in income statement, balance sheet, or other specific sections of the form
g_	Gross profit and loss items	Variables with this subscript are based on the gross profit and loss section of the income statement
k_	Assets	Asset variables in the balance sheet
l_	Liabilities	Liability variables reported in balance sheet
e_	Equity	Equity variables reported in balance sheet
y_	Income	Income variables reported in the balance sheet
x_	Expense	Expense variables reported in the balance sheet
irp5_	IRP5	Aggregated variables from the IRP5 firm-level dataset
vat_	VAT	Aggregated variables from the VAT firm-level dataset
cust_	Customs	Aggregated variables from the customs firm-level dataset
cgl_	Capital gain/loss	Capital gain/loss variables reported on the ITR14 form

Note: this table provides the descriptions of the variable prefixes and states the variable sources.

Source: authors' illustration.

Figure A1: Capital stock by source



Source: authors' illustration.

The most significant variable additions to the panel include variables from the transfer pricing and capital gains tax sections of the ITR14 form.

A1 New subsections and labelling

Although the variable names in the new panel remain unchanged, the labelling of each variable has changed slightly driven by the ambition to strengthen the clarity on where in the ITR14 tax form variables are derived from. The prefix of each variable label is now directly derived from the subsection in the ITR14 tax form to which the variable belongs. If the sub-section is 'international', the variable will be labelled as in Appendix Table A2.

Table A2: Variable labelling convention

Variable name	Primary section in the ITR14 tax form	Variable name with prefix	Sub-section in the ITR14 tax form	Label
fgnassinv	Company characteristics	*c_fgnassinv*	International	*International*—company owns foreign assets or investments

Source: authors' illustration.

Table A3: Customs firm-level variable name changes

Old dataset names	New dataset names
cust_exp_total	cust_export
cust_imp_total	cust_import
cust_imp_HS	cust_productimp_HS6
cust_imp_HS4	cust_productimp_HS4
cust_exp_HS6	cust_productexp_HS6
cust_exp_HS4	cust_productexp_HS4
cust_exp_countries	cust_countriesexp

Source: authors' illustration of name changes in the customs firm-level data v4.0 (National Treasury and UNU-WIDER 2021b).

Table A4: Labelling convention for a selection of customs variables

	Variable	Label
Raw variables	*AgentCode*	Customs-2.7 Agent code
	Tariff	Customs-5.5 Tariff code
Derived variables	*MainTrans*	Customs-derived Main transport used for trade
	MainAgent	Customs-derived Main agent used for trade

Source: authors' illustration based on customs transaction-level data v4.0 (National Treasury and UNU-WIDER 2021c).

A2 Recommended data citations

To encourage good research practice when using the administrative data at the National Treasury Secure Data Facility in Pretoria, we provide a list of recommended sources for the CIT-IRP5 panel as well as all the component datasets available to researchers in the data facility.

In-text example: National Treasury and UNU-WIDER (2021).

Reference list examples

Brink, D., and M. Kilumelume (2021). 'Deflator Variables Supplemental Data [dataset]. Version 1.0'. Pretoria: National Treasury and UNU-WIDER [distributor of the dataset], 2021.

Budlender, J., and A. Ebrahim (2019). 'Industry Variables Supplemental Data [dataset]. Version 1.0'. Pretoria: National Treasury and UNU-WIDER [distributor of the dataset], 2019.

National Treasury and UNU-WIDER (2021). 'CIT-IRP5 Firm-Level Panel 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2021.

National Treasury and UNU-WIDER (2020). 'CIT Firm-Level Panel 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'IRP5 Worker-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'IRP5 Firm-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'Customs Transaction-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

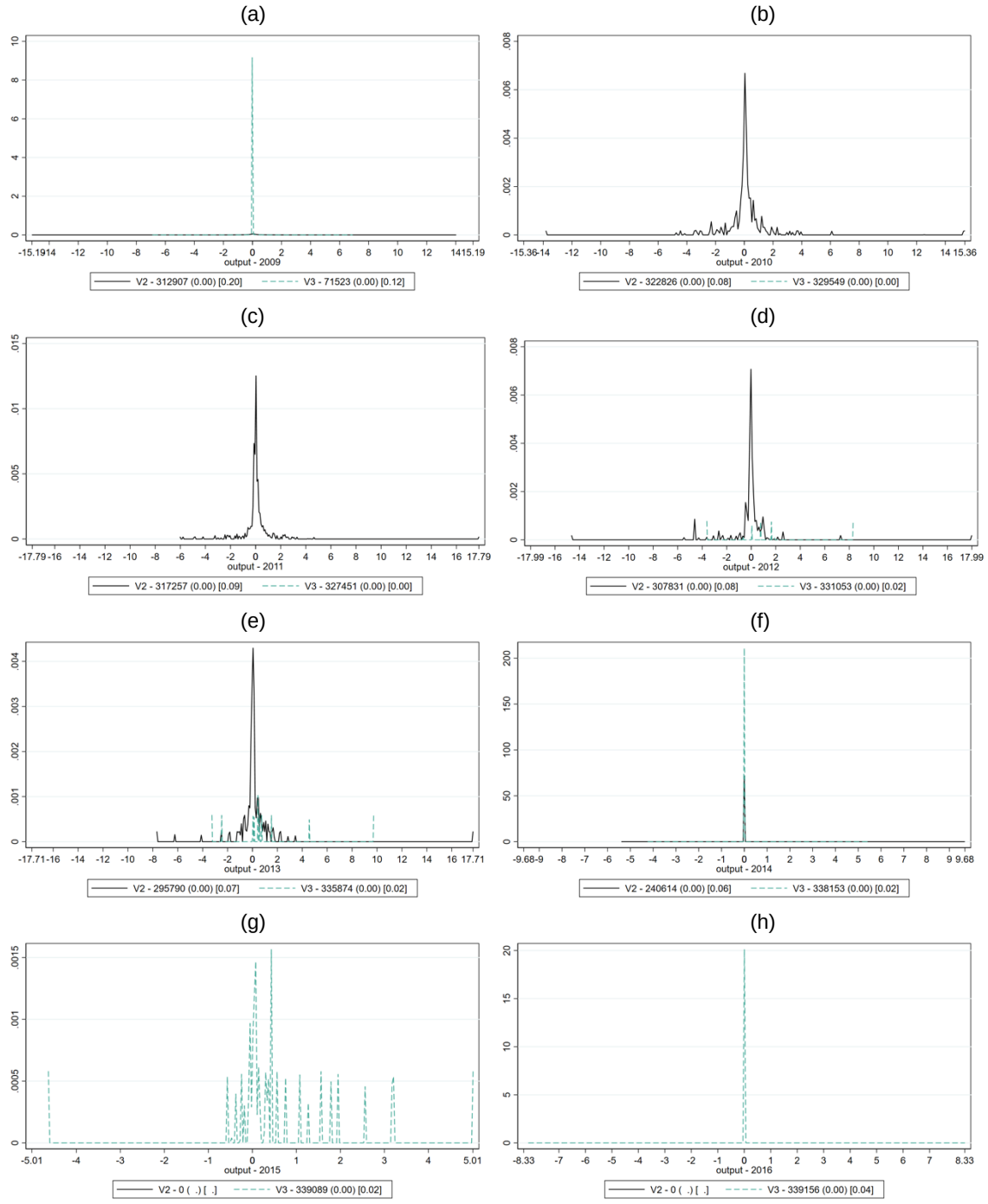
National Treasury and UNU-WIDER (2020). 'Customs Firm-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'VAT Firm-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

National Treasury and UNU-WIDER (2020). 'VAT Transaction-Level Data 2008–2018 [dataset]. Version 4.0'. Pretoria: South African Revenue Service [producer of the original data], 2019. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset], 2020.

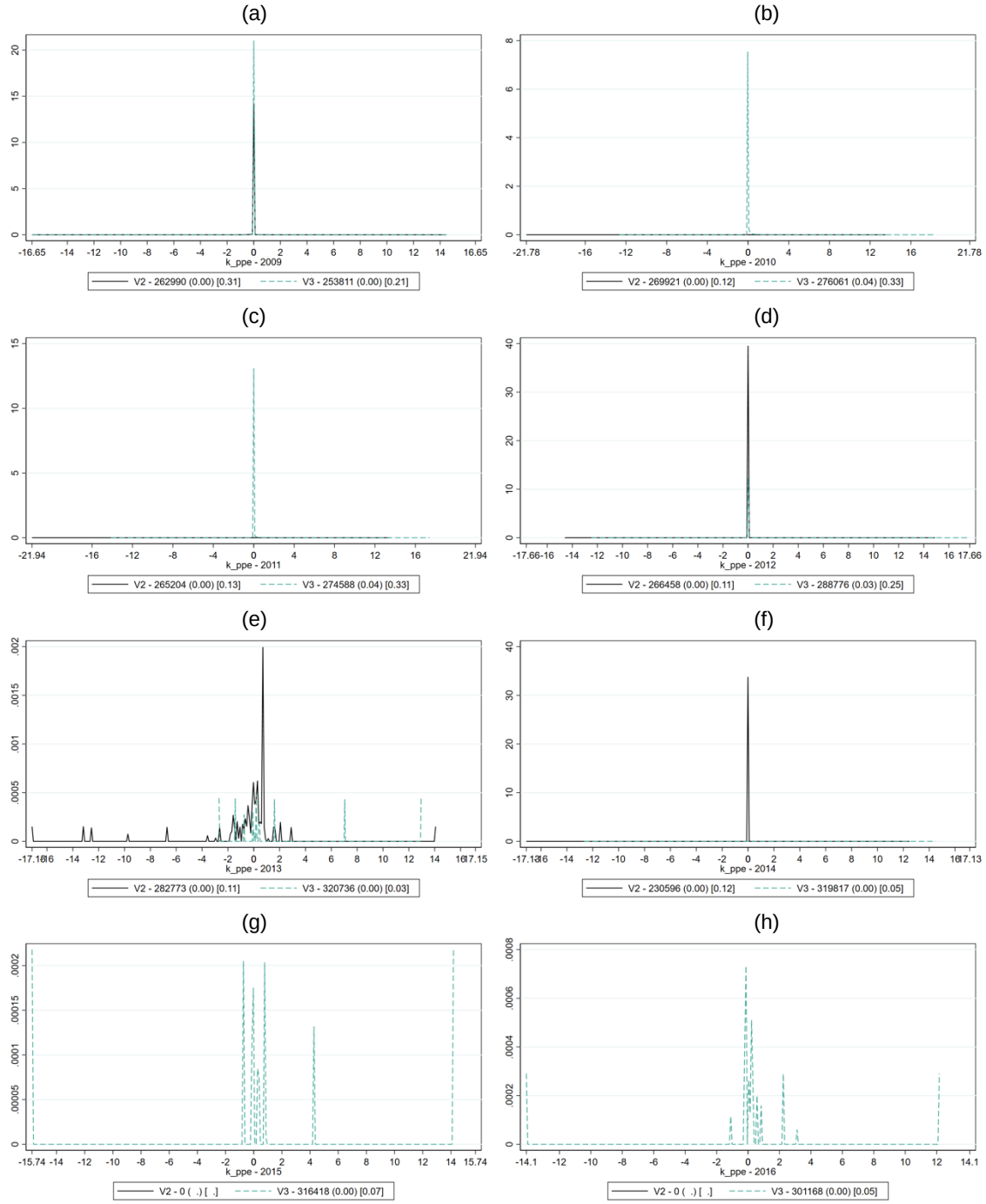
Appendix B

Figure B1: Distribution of differences in sales between panel versions by tax year



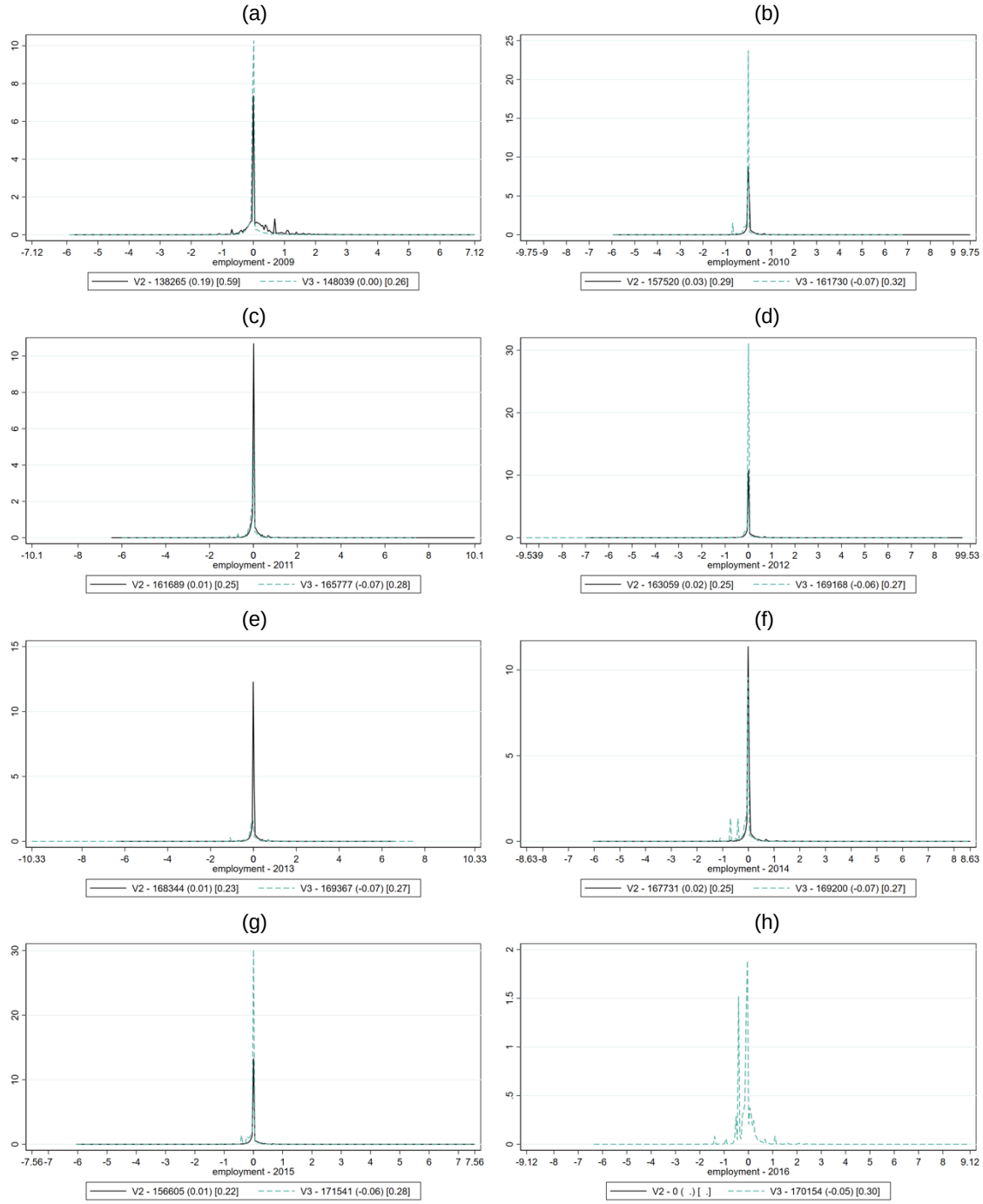
Source: authors' calculations based on CIT-IRP5 firm-level panel data v2.0, v3.0, and v4.0 (National Treasury and UNU-WIDER 2019, 2020, 2021a).

Figure B2: Distribution of differences in property, plant, and equipment data between panel versions by tax year



Source: authors' estimates using CIT-IRP5 firm-level panel data v2.0, v3.0, and v4.0 (National Treasury and UNU-WIDER 2019, 2020, 2021a).

Figure B3 Distribution of differences in weighted employment between panel versions by tax year



Source: authors' calculations based on CIT-IRP5 firm-level panel data v2.0, v3.0, and v4.0 (National Treasury and UNU-WIDER 2019, 2020, 2021a).