SA-TIED
Southern Africa – Towards Inclusive Economic Development

Simulating personal income tax in South Africa using administrative data and survey data

A comparison of PITMOD and SAMOD for tax year 2018

Wynnona Steyn, Alexius Sithole, Winile Ngobeni, Eva Muwanga-Zake, Helen Barnes, Michael Noble, David McLennan,Gemma Wright, and Katrin Gasior

SA-TIED Working Paper #186 | July 2021

UNITED NATIONS UNIVERSITY
UNU-WIDER

national treasury
Department:
National Treasury
REPUBLIC OF SOUTH AFRICA

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI

planning, monitoring & evaluation
Department:
Planning, Monitoring and Evaluation
REPUBLIC OF SOUTH AFRICA

the dtic
Department:
Trade, Industry and Competition
REPUBLIC OF SOUTH AFRICA

SARS

TIPS

About the project

Southern Africa –Towards Inclusive Economic Development (SA-TIED)

SA-TIED is a unique collaboration between local and international research institutes and the government of South Africa. Its primary goal is to improve the interface between research and policy by producing cutting-edge research for inclusive growth and economic transformation in the southern African region. It is hoped that the SA-TIED programme will lead to greater institutional and individual capacities, improve database management and data analysis, and provide research outputs that assist in the formulation of evidence-based economic policy.

The collaboration is between the United Nations University World Institute for Development Economics Research (UNU-WIDER), the National Treasury of South Africa, the International Food Policy Research Institute (IFPRI), the Department of Monitoring, Planning, and Evaluation, the Department of Trade and Industry, South African Revenue Services, Trade and Industrial Policy Strategies, and other universities and institutes. It is funded by the National Treasury of South Africa, the Department of Trade and Industry of South Africa, the Delegation of the European Union to South Africa, IFPRI, and UNU-WIDER through the Institute's contributions from Finland, Sweden, and the United Kingdom to its research programme.

Corresponding author: gemma.wright@saspri.org

The views expressed in this paper are those of the author(s), and do not necessarily reflect the views of the of the SA-TIED programme partners or its donors.

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2021/120

Simulating personal income tax in South Africa using administrative data and survey data

A comparison of PITMOD and SAMOD for tax year 2018

Wynnona Steyn,[1] Alexius Sithole,[1] Winile Ngobeni,[1] Eva Muwanga-Zake,[1] Helen Barnes,[2] Michael Noble,[2] David McLennan,[2] Gemma Wright,[2] and Katrin Gasior[2]

July 2021

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

Abstract: In this paper we explore South Africa's personal income tax system using two microsimulation models. The first, SAMOD, simulates personal income tax and social benefits using a dataset derived from the nationally representative National Income Dynamics Study survey. The second, PITMOD, simulates the personal income tax system and is underpinned by a dataset comprising a full extract of anonymized individual-level administrative tax data especially constructed for this purpose. The two models have a common framework in the form of the EUROMOD microsimulation software and interface, and have a common policy timepoint of 1 March 2017. Discrepancies between the simulated personal income tax generated by each model are explored in order to better understand the strengths and weaknesses of the two models.

Key words: microsimulation, personal income tax, income distribution, South Africa

JEL classification: C63, C81, D31, H24

Acknowledgements: UNU-WIDER is thanked for supporting the construction of PITMOD as part of the SA-TIED programme of work. The University of Essex is thanked for granting a licence to SASPRI and SARS to use EUROMOD. The European Commission is acknowledged for supporting the development of EUROMOD. The authors would additionally like to thank the following people for their contributions: Rebone Gcabo, Mamiky Leolo, Caroline Ramara, Maria Malope, Nombulelo Ngcamu, Lucia Sithole, Alet Wasmuth, Lilian Khumalo, Mojalefa Molapo, Aleseng Moshoette, Khanyisa Phika, and Dinah Sekhuthe (SARS); Professor Ada Jansen (University of Stellenbosch, RSA); Dr Chrysa Leventi (Ministry of Finance, Greece); and Kostas Manios (University of Essex, UK).

[1] South African Revenue Service, National Treasury, Government of South Africa, Pretoria, South Africa; [2] Southern African Social Policy Research Insights, Hove, United Kingdom; corresponding author: gemma.wright@saspri.org

This study has been prepared within the UNU- 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 978-92-9267-060-3

https://doi.org/10.35188/UNU-WIDER/2021/060-3

Typescript prepared by Joseph Laredo.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

The views expressed in this paper are those of the author(s), and do not necessarily reflect the views of the Institute or the United Nations University, nor the programme/project donors.

1 Introduction

This Working Paper presents the main findings from a recent study that explored South Africa's personal income tax (PIT) system using two microsimulation models. The first model, SAMOD, simulates personal income tax and social benefits using a dataset derived from the fifth wave of the nationally representative National Income Dynamics Study (NIDS) survey. The second model, PITMOD, simulates the PIT system only and is underpinned by a dataset comprising a full extract of anonymized individual-level administrative tax data especially constructed for this purpose.

Tax–benefit microsimulation models are important tools for policy analysis and, although they are usually underpinned by survey datasets, there is growing interest in the use of administrative datasets for modelling (e.g. Figari et al. 2014). Access to South African administrative data in general for research and policy purposes is on the increase (McLennan et al. 2017). Several studies have already used administrative data on income tax; for example, Orthofer (2016) and Hundenborn et al. (2017 and 2019) used income tax data to validate the income data in NIDS, and Bassier and Woolard (2020) and Wittenberg (2017) used income tax data to validate the income data in the Post-Apartheid Labour Market Series (PALMS) dataset. Usage will continue to increase following the recent construction of a South African Revenue Service National Treasury (SARS-NT) Individual Panel of taxpayers (Arndt 2018; Ebrahim and Axelson 2019).

For the purposes of microsimulation, the quality of the income variables in the underpinning dataset—whether derived from a survey or from administrative data—is critically important and will have a direct impact on the simulation of taxes including PIT (e.g. Ceriani et al. 2013; Sutherland 2018). This paper builds on an earlier study that compared the simulation of PIT in South Africa using two social surveys—the Living Conditions Survey (LCS) 2014/15 (Stats SA 2017a) and NIDS Wave 4 Version 1.1 (SALDRU 2014)—as underpinning datasets for SAMOD (Wright et al. 2018). It was found that the numbers of taxpayers identified in SAMOD using LCS and NIDS data were similar to each other and corresponded closely to the published figure for assessed individuals. However, SAMOD simulated less PIT revenue using the LCS dataset than with NIDS, and both surveys simulated less PIT revenue than published figures from administrative data. It was found that both surveys suffer from unit missing data for high-income individuals (R1 million and over), with SAMOD simulating far fewer taxpayers in the highest taxable income group, whether using the LCS or NIDS, than reported by the National Treasury (2015). This finding was in line with the assessments of NIDS income data by Hundenborn et al. (2017) and Rasmussen (2017). It is therefore particularly important to be able to harness administrative data for the upper part of the income distribution (see also Shine et al. 2019).

In terms of the structure of the paper, Section 2 provides a brief account of South Africa's PIT policy. Section 3 introduces the two models, SAMOD and PITMOD, with respect to their underpinning datasets and how they each simulate the PIT policy system. Section 4 compares the distribution of income data in the SAMOD and PITMOD datasets; presents SAMOD and PITMOD PIT simulations for the tax year 2017/18 and compares these with published administrative aggregate data on PIT; and compares SAMOD and PITMOD simulations of PIT by taxable income group and by income tax band. Section 5 concludes with a discussion of the findings and sets out a roadmap for providing researchers with access to PITMOD.

2 The South African personal income tax system in context

PIT serves the purpose of raising revenue and ensuring that equity objectives are reached. Since high-income earners generally benefit most from globalization, the PIT system is ideally suited to capturing revenue from these income groups for redistributive purposes (Tanzi 2004). Tax policy entails a government deciding what taxes to levy, in what amounts, and on whom (OECD 2013). In designing or reforming a tax system, aspects such as fairness and equity (i.e. who pays the tax and how it affects the distibution of income) and allocative efficiency (i.e. the possible distortionary effects of taxes on economic activity) are key considerations. A tax is progressive when the average tax rate increases as income increases (Mirrlees et al. 2011). The progressivity of the income tax system is influenced by the structure of the system (i.e. the minimum tax threshold and the progression of the marginal tax rates) (Mirrlees et al. 2011), as well as the taxable income base (i.e. the taxation of the different sources of income, such as wages versus capital income and the deductions allowed for specific social and economic goals). According to Mirrlees et al. (2011), public spending requires taxation, which is not costless given its economic impact. Hence, the challenge in desiging a tax system is achieving social and economic objectives while limiting the welfare-reducing side-effects. Policy-makers should also ensure that tax compliance and tax administration costs are minimized. The optimal system and the required reform measures are dependent on the need to improve tax revenue mobilization and the tax administrative capacity of a country. It is important to keep the system simple and transparent whilst adhering to the principles of a good tax system such as neutrality, fairness, and certainty (Tanzi 2001).

Tax reform options are diverse, but globalization has directed tax reform proposals towards conformity. Proposals for adjustments to the structure of the PIT system include changes to marginal tax rates and thresholds, while tax base reforms include changes to exempt income, the distinction between capital income and labour income, and tax deductions to simulate the desired revenue and distributional impact. As pressure on growth and on public finances mounts, the need for an efficient and effective tax system is greater than ever.

The pressure on fiscal resouces is a stark reality in South Africa, a middle-income developing country with a highly skewed income distribution (Statistics South Africa 2017b). The current need to reduce poverty levels requires additional public expenditure and therefore higher tax revenue. These opposing pressures force policy-makers to reduce inefficiencies in public expenditure and at the same time to align the tax system to generate sufficient revenue in an equitable manner (Steenekamp 2012). According to Tanzi (2004: 534), as cited by Steenekamp (2012), the two 'work horses' that must carry this burden are value-added tax (VAT) and PIT. VAT was introduced in South Africa in 1991 and has become what is generally considered to be an efficient revenue source but a regressive tax in terms of income. PIT is the most important source of tax revenue in South Africa, contributing 38 per cent of total tax revenue in 2017/18 (National Treasury and SARS 2020: 19).

The PIT system is broad-based with few exemptions and deductions. There are seven income tax brackets and marginal tax rates that vary from 18 per cent for the lowest income bracket to 45 per cent for the upper bracket—for taxable income above ZAR1.5 million per annum. The minimum tax threshold (currently ZAR350,000) is determined by the marginal tax rate and the applicable tax credits (or rebates), namely a primary tax credit for persons under the age of 65 years, plus a secondary tax credit for persons 65 years and older and a tertiary tax credit for persons older than 75 years. The minimum tax threshold of ZAR75,750 for persons under the age of 65 years for the

2017/18 fiscal year was less than the GDP per capita for 2018 (R83,853.65); and the maximum income tax bracket of ZAR1.5 million was close to 18 times the GDP per capita amount.[1]

In 2001, South Africa adopted the residence basis of taxation and the taxing of capital gains. Capital gains are not taxed separately and are not indexed for inflation, but to provide relief for inflationary gains only a portion of realized gains are included as part of taxable income, referred to as the inclusion rate. The PIT system in South Africa therefore conforms to a semi-comprehensive tax system, as is evident in the taxation of labour income versus capital income—for instance interest, capital gains, and profits from personal businesses. The principle of neutrality is sacrificed but the after-tax Gini coefficient can be improved. The comprehensive PIT system is open to tax arbitrage, whereby individuals restructure their tax affairs to minimize their tax burden. Tax revenue may be compromised, as well as the principles of vertical and horizontal equity. In addition, compliance gaps in terms of accurate declarations in a self-assessment system tend to be higher.

Given the current thinking and empirical evidence on PIT policy and reform, the direction of fundamental PIT reform globally can be categorized into four alternative PIT approaches (Steenekamp 2012):

- a flat tax system, which applies a single tax rate with a basic tax allowance;
- a dual income tax system, which combines a single (low) tax rate on capital income with a progressive rate structure for labour income;
- a comprehensive PIT system, which combines a progressive rate schedule for all sources of income with a system of tax reliefs;
- presumptive taxation, which is an administrative assessment system using indicators such as assets or turnover.

The South African tax system has undergone reforms in many respects since the democratization of the country in 1994. The reforms have aimed to improve revenue collection efficiency, broaden the tax base, and adjust the minimum tax thresholds and income brackets for inflation on an annual basis. In 2013, the Minister of Finance appointed a tax review committee, tasked with analysing the tax system and considering reform options (National Treasury 2014). At the time, the Ministry of Finance indicated that the relatively modest economic growth and profound social challenges in South Africa (such as poverty, inequality, and persistent unemployment) necessitated a review of the role of the tax system as part of a well functioning and coherent fiscal planning framework. Furthermore, it was felt that the progressivity of the South African tax system could contribute to the social objectives of building a cohesive and inclusive society by raising revenue to effect redistribution (Ministry of Finance 2013).

3 Methodology

The findings presented in this paper draw from the results of two microsimulation models, SAMOD and PITMOD. SAMOD is a South African tax–benefit microsimulation model which has been developed for use by the government over the past decade (e.g. Wilkinson 2009; Wright et al. 2011). PITMOD, a model dedicated to the simulation of South Africa's PIT policy, was developed for internal use by the South African Revenue Service (SARS) and the National Treasury (NT), as well as ultimately for broader academic research purposes.

[1] The income per capita for 2018 in current terms is ZAR58,732 (IHS Markit).

SAMOD and PITMOD have a common framework in the form of the EUROMOD microsimulation software and interface (University of Essex 2020), which has the advantage of methodological and conceptual harmony (e.g. Sutherland 2001; Sutherland and Figari 2013; Tammik 2018).[2] SAMOD represents the first attempt to use the EUROMOD microsimulation software in a developing country context. PITMOD is similarly the first attempt to use EUROMOD software for a model underpinned by administrative data in a developing country context, although there are examples in higher-income countries, for example, Greece (Leventi et al. 2013) and Belgium (Verbist and Mechelen 2020). Examples of PIT models that are underpinned wholly or partially by administrative data (but which do not use the EUROMOD software) are those of France (Jelloul et al. 2019), Germany (Flory and Stöwhase 2012), Italy (Miola and Manzo 2021), and Spain (Bover et al. 2017).

The models also use a common *policy* timepoint that eliminates the extent to which differences observed between the two models could be due to policy changes over time. The South African tax year runs from 1 March to 28/29 February. Both SAMOD and PITMOD simulate the PIT rules applicable at 1 March 2017, which is referred to by SARS as the start of the 2018 tax year.

The relative strengths and weaknesses of the two models (summarized in Table 1) are as follows: SAMOD applies the main tax and benefit policies applicable to individuals to a nationally representative survey dataset, which enables results to be generated for the entire income distribution of South Africa, not just for taxpayers. In contrast, PITMOD focuses exclusively on the PIT system and does not simulate any of the other taxes and benefits. Since PITMOD is underpinned by the administrative data records of all compliant taxpayers, any findings using PITMOD are restricted to that subset of the population, which will exclude non-compliant taxpayers as well as individuals who are not on the SARS registry for other reasons. However, this is also one of PITMOD's main strengths, as it thereby captures the upper part of the income distribution better than is possible using survey data.

Table 1: Summary of SAMOD's and PITMOD's main strengths and weaknesses

	SAMOD	PITMOD
Provides results for the entire income distribution of South Africa	✓	✗
Simulates the main tax and benefit arrangements that apply to individuals	✓	✗
Simulates PIT policy in fine detail, allowing for extensive policy reform explorations	✗	✓
Uses an underpinning dataset comprising all compliant taxpayers	✗	✓

Source: authors' construction.

In the rest of this section, the simulation of PIT is elaborated, both in terms of how comprehensively the PIT policy is simulated by each model (Section 3.1) and in terms of the construction of the underpinning datasets used by the models (Section 3.2).

3.1 Simulating the personal income tax policy in SAMOD and PITMOD

South Africa's tax and benefit policy rules are set out in SAMOD as modules or distinct 'policies' (in EUROMOD terminology), alongside more general rules relating to the framework of the model or for the purposes of later analysis of the model output via the Statistics Presenter tool. In SAMOD, the full tax–benefit system—in so far as it is possible to model with survey data—is presented. PIT is only partially simulated within SAMOD (see Tables 2, 4, and 5) and the PIT rules

[2] See https://www.euromod.ac.uk/about/what-is-euromod.

are condensed into four modules or 'policies'.[3] In contrast, in PITMOD, only the rules relating to PIT are presented, and there are fewer general rules relating to the framework of the model and analysis. Using administrative data as the underpinning dataset for PITMOD, it is possible to model almost all elements of the PIT policy rules; and most of these have their own 'policy' in PITMOD.[4]

When the SAMOD and PITMOD models are run, every individual in the underpinning dataset is tested against the full spectrum of policy rules. In PITMOD, tax liability is calculated as the last step in the model.[5]

The modules or policies relating to PIT within SAMOD and PITMOD can be considered as comprising five separate groups relating to income, deductions, tax credits, lump sums, and tax liability (amount of PIT payable). The first two—income and deductions—are used to calculate final taxable income, which is then used in the calculations of tax credits and tax liability. The group that relates to lump sum income intersects with the other four groups only at the stage at which final tax liability is calculated. These five groupings are discussed in brief next. The focus is on PITMOD rather than SAMOD, although SAMOD is compared with PITMOD in many of the tables. Further details about SAMOD can be found in Wright et al. (2018).

Sources of taxable income

Various sources of income are included in the taxable income concept in PITMOD. Some have exempt amounts or exclusions, and some have their own tax schedules. In other words, some income sources need a separate policy within PITMOD to calculate the amount of income to be added to taxable income, while most of the other income sources go straight into the calculation of taxable income used towards the end of the model. Income from lump sums (retirement, severance, and withdrawal) is taxed separately. Table 2 summarizes the different sources of income that are taken into account in the two models.

[3] The four policies calculate rebates, medical tax credits, income tax on lump sums, and final income tax.

[4] The PIT policy rules were obtained from the following SARS resources:
IT-AE-36-G05 - Comprehensive Guide to the Income Tax Return for Individuals - External Guide
LAPD-IT-G01 - Guide on Income Tax and the Individual
PAYE-GEN-01-G03 - Guide for Employers in respect of Allowances - External Guide
PAYE-AE-06-G06 - Guide for Codes Applicable to Employees' Tax Certificates 2019 - External Guide
LAPD-IT-G03 - Guide on the Calculation of the Tax Payable on Lump Sum Benefits
LAPD-IT-G19 - Comprehensive Guide to Dividends Tax
LAPD-LPrep-Draft-2019-75 - Draft IN 18 Issue 4 - Rebates and Deductions for Foreign Taxes on Income
LAPD-CGT-G02 - The ABC of Capital Gains Tax for Individuals
LAPD-IT-G07 - Guide on the Determination of Medical Tax Credits
https://www.sars.gov.za/types-of-tax/personal-income-tax/

[5] This means in PITMOD that for the most straightforward case of an individual who only has employee income that meets the criteria for not submitting a tax return, only the main income tax, rebates, and final tax liability calculations will be performed, although the individual will still 'pass through' the intervening policies without any calculations being performed (as the relevant income or expenditure is not present).

Table 2: Summary of sources of income taken into account in the simulation of PIT in SAMOD and PITMOD

INCOME	Included in SAMOD	Included in PITMOD
Employee income: Salaries and wages Fringe benefits Allowances Overtime Options to purchase shares Pensions Bonuses Restraint of trade Annuities Director fees Incentive awards Commission	Some income sources, dependent on information in survey data	Y, all income sources are available in the administrative data
Business income	Y	Y
Farming income	Y	Y
Investment income:		
Interest—exemption (amounts vary by age and whether local or foreign interest income)	Y	Y
Foreign dividends – exemption	N	Y
Other dividends (REIT and deemed)	N	Y
Capital gains/losses – annual exclusion and net capital gain multiplied by inclusion rate	N	Y
Local rental income profit/loss	Y	Y
Other taxable income: Royalties profit/loss Other	Some income sources, dependent on information in survey data	Y

Source: authors' compilation.

Although the main use of taxable income is in calculating tax liability, it is required in different forms in some of the deduction and tax credit policies and is therefore discussed next.

Taxable income is calculated by subtracting deductions from relevant income sources. This calculation is made via a dedicated type of module or 'function' within PITMOD called an 'income list'. Technically an income list is the aggregate of several variables, which are added or subtracted to build the aggregate. Three definitions of taxable income—and therefore three income lists—are required in PITMOD (see Table 3).[6] Two are used for calculating the deduction for retirement fund contributions, and the third is used in the calculation of medical tax credits and main income tax liability (as distinct from tax liability from lump sums—see below). The main differences between the three definitions are the inclusion or exclusion of capital gains (included in two of the three income lists) and whether taxable income is income before or after deductions. Lump sums are excluded from all three income lists in PITMOD as they are dealt with separately.

[6] Other income lists are used in PITMOD, including employee income (il_employee), remuneration (il_remuneration), and deductions (il_general_deductions).

Table 3: Summary of PITMOD's income lists for taxable income

Income source	il_taxabley01	il_taxabley02	il_taxabley03
	Used in calculating the deduction for retirement fund contributions (Part b(ii))	Used in calculating the deduction for retirement fund contributions (Part c)	Used in the calculation of medical tax credits and main income tax liability
Employment income	Y	Y	Y
Passive income	Y	Y	Y
Taxable capital gains	Y	N	Y
Deductions	N	N	Y

Source: authors' compilation.

Finally, certain income, including investment income and capital gains, is shared equally between the taxpayer and their spouse if they are married in community of property, and so an individual is taxed on half of their own income and half of their spouse's income. Any exemptions apply to each taxpayer. All other taxable income is deemed to be the income of the taxpayer who receives the income and forms part of their taxable income only. The rules regarding marriage in community of property are taken into account in PITMOD, with a 50 per cent split applied to income from interest and dividends (total income in each category for the couple is declared on the tax return) for anyone married in community of property (flagged by the variable *dms* in the administrative dataset). For income from rental and capital gains the situation is slightly different: although a 50 per cent split should be applied, the split is already applied to the data that are brought into the model so that it is not necessary to undertake the 50 per cent split step on the model.

Deductions

With regard to deductions, a number of deductions from taxable income need to be made, where applicable, before tax liability is calculated. Deductions therefore reduce a taxpayer's taxable income. In PITMOD, the deduction for retirement contributions is in a separate policy. Most of the other deductions are straightforward and do not require a separate policy in PITMOD for a calculation to be performed. These deductions are therefore included either in the employee income policy (deductions relating to expenses against allowances or against commission income) or in a deductions policy (all other deductions, including donations and VCC investments).[7] Table 4 summarizes the extent to which PIT deductions are simulated in SAMOD and PITMOD.

[7] Within this deductions policy, the deduction for allowable accountancy/administration expenses has to be treated separately as the deduction can only be made if the taxpayer has certain types of income and cannot be made against salary or wage income. The relevant income is included in another income list called il_accountancy_income.

Table 4: Summary of the extent to which PIT deductions are simulated in SAMOD and PITMOD

Deductions	Included in SAMOD	Included in PITMOD
Retirement contributions	Y	Y
Other deductions:	N	Y
Donations		
Travel claim against travel allowance		
Employer-provided vehicle (operating lease or other arrangement)		
Expenses against local and/or foreign taxable subsistence allowance		
Depreciation		
Home office expenses		
Travel expenses (no allowance)		
Amounts refunded		
Allowable accountancy/administration expenses		
Legal costs		
Bad and doubtful debts		
Section 8C losses		
Assumed expenses of holders of public office		
Remuneration taxed on IRP5 but complying with exemption in terms of Section 10(1)(o)(i) or (ii)		
Commission income expenditure		
Investments in venture capital companies		

Source: authors' compilation.

Tax credits

Tax credits reduce a taxpayer's tax liability. Three different types of tax credit need to be taken into account: rebates, medical tax credits, and foreign tax credits. Only the first two types of tax credit are modelled in PITMOD, each in a separate policy. It is intended that the foreign tax credit policy will be modelled in the next phase. Table 5 summarizes the extent to which tax credits are simulated in SAMOD and PITMOD.

Table 5: Summary of the extent to which tax credits are simulated in SAMOD and PITMOD

Tax credits	Included in SAMOD	Included in PITMOD
Rebates	Y	Y
Medical tax credits	Y	Y
Foreign tax credits	N (not possible)	N (may be possible in future)

Source: Authors' compilation.

Tax on lump sums

Individuals are required to pay tax on three different types of lump sum income: retirement,[8] severance,[9] and withdrawal.[10] Income from lump sums is taxed separately and each type of lump sum has its own tax schedule. Tax on lump sums is not currently simulated in PITMOD as it is not possible to identify appropriate source codes to enable the policy to be modelled accurately. It was therefore decided to use the variable from the administrative data relating to tax paid on lump sums. The infrastructure for a lump sums policy is included in PITMOD with a view to modelling

[8] A retirement benefit is a lump sum paid to a member of a retirement fund as a result of death, or when the member reaches retirement age or is retrenched.

[9] A severance benefit refers to a lump sum paid to a person from their employer as a result of retrenchment. The amounts paid by an employer are unrelated to amounts paid by the person's retirement fund.

[10] A withdrawal benefit is an amount paid to a member of a retirement fund when they terminate their membership in that retirement fund before reaching retirement age, for any reason other than retirement, death, or retrenchment.

tax on lump sums in the next phase. The tax liability from lump sums is added to the main income tax liability as the final step of calculating overall income tax liability.

Tax liability

As discussed above, taxable income used in the calculation of the main income tax liability (il_taxabley03) includes employment income, passive income, taxable capital gains, and deductions.

As seen above, the income sources that are taken into account in SAMOD and PITMOD differ to a certain extent (Table 2), as do the deductions (Table 4) that are taken from taxable income to arrive at the final taxable income amount to be used in the calculation of tax liability.

In order to calculate tax liability, a tax schedule comprising seven tax brackets is applied to taxable income. Then the tax credits calculated on the model are subtracted from the tax liability, and tax on lump sums is added to the main tax liability. The calculation of tax liability in PITMOD (and SAMOD) can therefore be summarized as follows:

Tax liability = (tax payable on taxable income – rebates – medical tax credits (with a lower limit of zero)) + tax payable on lump sums

Using the EUROMOD nomenclature of the PITMOD (and SAMOD) variables, the calculation of tax liability, tin_s, in PITMOD can be summarized as follows:

tin_s = (tax payable on il_taxabley03 (or tingt_s) – tinta_s – tintchl_s (with a lower limit of zero)) + tinlu_s[11]

In summary, the modelling of PIT in SAMOD is partial as a result of the availability of information relating to PIT in the survey data, whereas the modelling of PIT in PITMOD is more comprehensive (currently only foreign tax credits and tax on lump sums are not simulated in PITMOD).

3.2 The underpinning datasets in SAMOD and PITMOD

SAMOD's underpinning dataset

SAMOD Version 7.4 is underpinned by a modified version of the fifth wave of the National Income Dynamics Study (NIDS) (SALDRU 2018). NIDS is a national panel study carried out by the University of Cape Town. Although it is designed as a panel study, a specific set of weights enable the dataset to be used as a cross-sectional, nationally representative dataset (Branson and Wittenberg 2019). All five waves of NIDS have been used as underpinning datasets for SAMOD (e.g. Wright et al. 2011).

The data for the fifth wave were collected between February 2017 and December 2017 with monetary values deflated to March 2017 by the NIDS team. These were then inflated to a timepoint of 30 June 2017 using the Consumer Price Index (Stats SA 2017c) as part of the process of preparing the dataset prior to importing it into SAMOD as the input dataset. The final dataset contains 10,659 households comprising 39,434 individuals. This is fewer cases than in the original NIDS data, where there are 10,842 successfully interviewed households comprising 40,944 individuals. Each of these individuals was given a weight if the household was successfully

[11] Tax payable on lump sums is called tinkt_s in SAMOD.

interviewed, regardless of whether the individual was successfully interviewed. Individuals who were not successfully interviewed could not be included in the SAMOD input dataset as almost all information was missing. The weights supplied in the NIDS dataset were therefore recalibrated to Statistics South Africa's population estimates for June 2017 to adjust for the cases that were dropped. The technique of iterative proportional fitting (IPF) (also referred to as 'raking') was used to adjust the weights.[12]

The variables that are used for the simulation of PIT in SAMOD are set out in Appendix A (see Table A1).

PITMOD's underpinning dataset

PITMOD is underpinned by an input dataset derived from a source administrative dataset compiled and supplied by SARS. The dataset is a composite of tax-specific administrative data held by SARS and external data on medical insurance scheme contributions, which are routinely supplied to SARS by third-party organizations. The dataset is a full extract and has been anonymized in-house by SARS.

The information in the dataset is mainly derived from a central enterprise data warehouse, allowing information from the IRP5/IT3a and ITR12 forms to be combined into one dataset. IRP5/IT3a is the employee tax certificate submitted by the employer on behalf of the employee. The IT3a element relates to people with a wage/salary but where no tax is deductible. ITR12 is the PIT return for individuals with employee income over the ZAR350,000 threshold, or individuals who work for more than one employer during the given tax year, or individuals with additional income or tax-related deductions and rebates not taken into account in the IRP5/IT3a employer return, or taxpayers who are not employees and therefore not part of the Pay-As-You-Earn (PAYE) system. Additional information on investment income is contained within the IT3b dataset as well as in the ITR12 return. These SARS data sources were supplemented with third-party information on medical insurance scheme contributions.

There is a distinction between 'returned data', which relates to the information contained within the IRP5/IT3a and ITR12 forms submitted to SARS, and 'assessed data', which relates to the final values calculated by SARS during the assessment process and which may or may not be different from the values in the returned data following revisions to the returned data during assessment.

An individual can be in both the IRP5/IT3a and ITR12 datasets, and there can be more than one record per taxpayer (e.g. if an individual moves from one job to another during the tax year, or has several concurrent jobs). However, this has been dealt with in the production of the dataset with a single overall composite IRP5/IT3a return compiled for each individual who had multiple IRP5/IT3a forms, and information from ITR12 and third-party medical insurance scheme data merged in-house using the taxpayer's unique Tax Reference Number and/or unique ID number. The dataset therefore has only one record per individual.

PITMOD's underpinning dataset contains 14.7 million individuals. The input variables that are used for the simulation of PIT in PITMOD are described in Appendix B alongside further details about the construction of the PITMOD dataset. Much of the model development was undertaken with a 10 per cent (or smaller) sample due to the difficulties of processing such a large dataset.

[12] The Stata .ado file 'ipfraking' was utilized.

The current version of PITMOD is underpinned by administrative tax records spanning the 2017/18 tax year (referred to as the 2018 year of assessment), i.e. data from 1 March 2017 to 28 February 2018. It is an optimal dataset for comparison with SAMOD Version 7.4 as the NIDS Wave 5 data collection period overlaps considerably with the relevant PITMOD administrative data collection period.

The three main differences between the PITMOD dataset and the SARS-NT Individual Panel of taxpayers (Ebrahim and Axelson 2019) are that the PITMOD dataset relates to one tax year, it contains just one record per individual, and it contains many more variables than the Panel so as to enable both the simulation of each of the elements of the PIT system and the simulation of potential PIT reforms.

4 Results

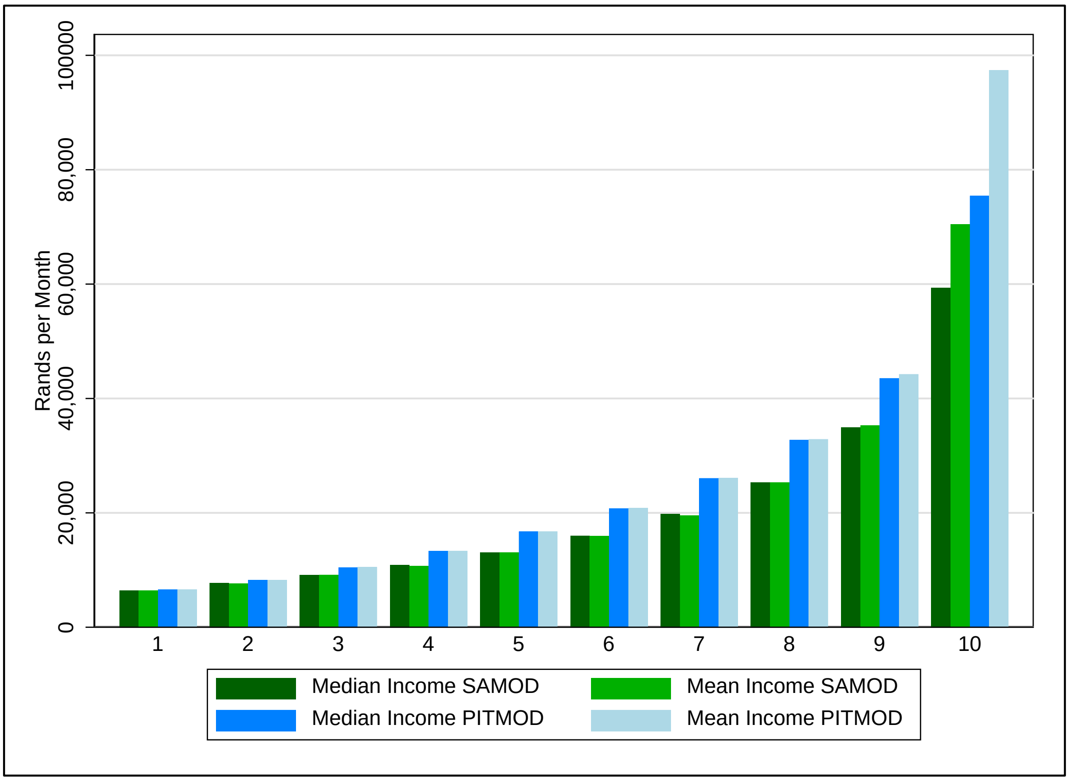
4.1 Comparing the income data in the underpinning datasets of SAMOD and PITMOD

Before presenting the simulated PIT amounts using SAMOD and PITMOD, it is instructive to compare pre-tax income. In an ideal world we would compare market income, but PITMOD's underpinning dataset contains only data on potentially taxable income. So, for example, sources of income such as private transfers (e.g. remittances), whilst part of market income, are not collected by SARS. Moreover, some sources of potentially taxable income, such as capital gains, are not available through the NIDS survey data underpinning SAMOD and are therefore not comparable. Certain sources of income such as rental income and income from investment are relatively rare in the NIDS dataset and again make comparisons challenging. We have therefore restricted the analysis in this section to two sources of income that can be meaningfully compared: employment income and taxable income (excluding capital gains).

For each dataset the mean and median income by decile of that particular type of income was calculated. The deciles derive from the respective data sources and, in the case of SAMOD, are weighted.

Employment income comprises employed earner income and self-employment income. In South Africa a PAYE system is in place for employed workers. Unless they have other special sources of income, those whose earnings are below ZAR350,000 per annum are not required to submit tax returns. However, the self-employed and those with higher earnings are so required. Therefore, the administrative data comprising employment income is partly derived from the IRP5 data and partly from the ITR12 data. In both SAMOD and PITMOD terminology this is expressed in the variables *yem* plus *yse*. The following chart shows the mean and median employment income in tax year 2017/18 by employment income decile.

Figure 1: Mean and median employment income in 2017/18 by decile, SAMOD and PITMOD

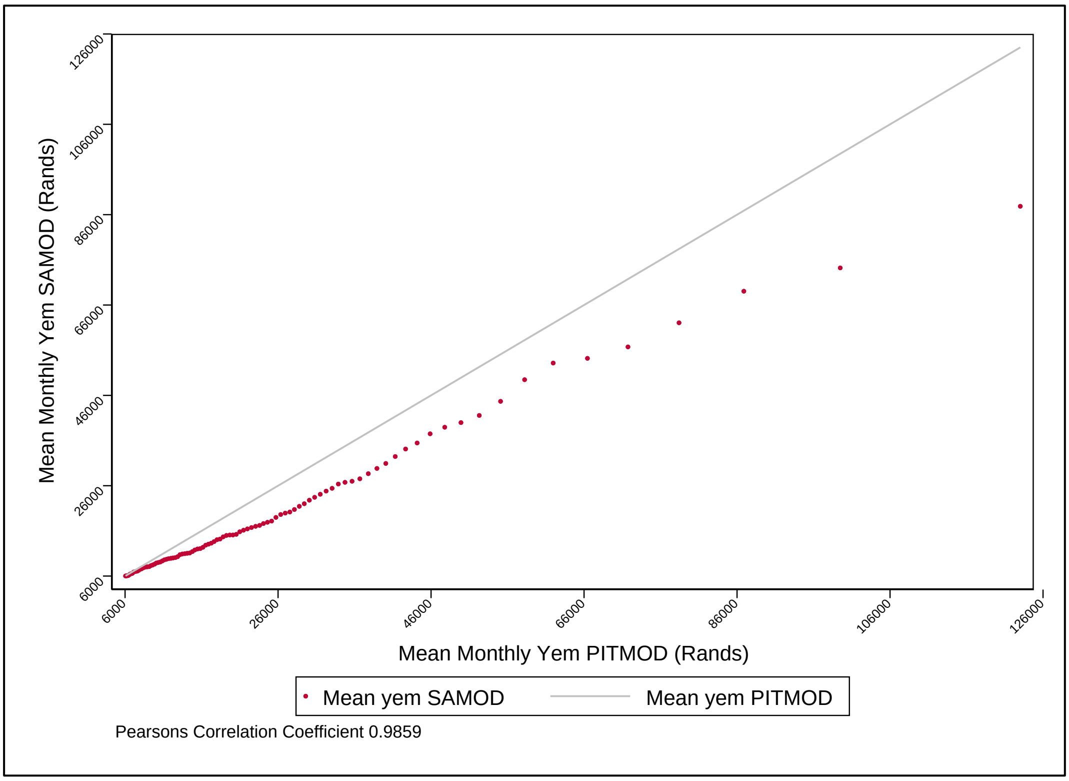


Note: the SAMOD and PITMOD cases include only those with a taxable income greater than zero.

Source: authors' calculations using SAMOD Version 7.4 (using NIDS Wave 5 Version 1.1) and PITMOD's underpinning administrative dataset.

It is evident from Figure 1 that the mean (or median) employment income using SAMOD trails behind PITMOD in all deciles, not just decile 10. The main component of employment income is employed earner income (*yem*). It is possible to look at the relationship between *yem* in the input dataset for SAMOD and the equivalent for PITMOD by calculating mean (and median) income for each percentile of the earned income distribution for each of the two models. These can then be scattered. Figure 2 shows the result for mean income for the centiles 1–99 (centile 100 is omitted because it is an extreme outlier).

Figure 2: Mean gross earned income (yem) for each percentile 1–99, SAMOD and PITMOD



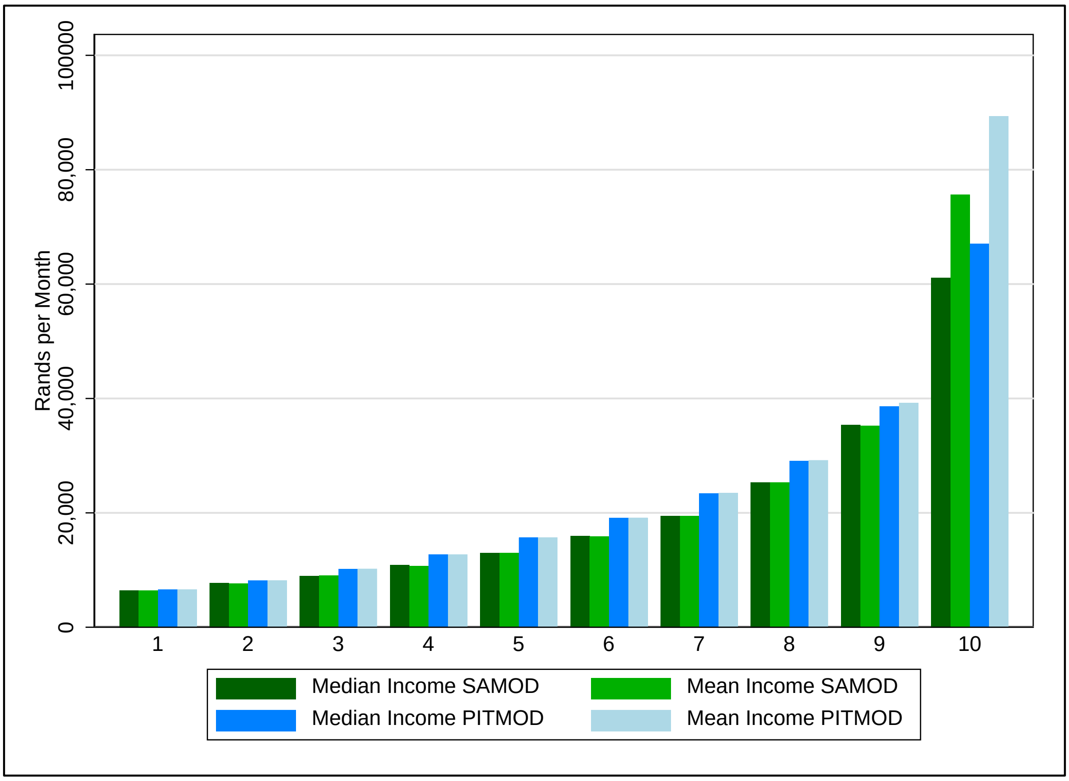
Note: the SAMOD and PITMOD cases include only those with a taxable earned income greater than ZAR6,000 per month.

Source: authors' calculations using SAMOD Version 7.4 (using NIDS Wave 5 Version 1.1) and PITMOD's underpinning administrative dataset.

One possible explanation for this relates to the NIDS data underpinning SAMOD. SAMOD requires the input dataset to contain a variable representing gross earnings. The NIDS team at the University of Cape Town undertake extensive cleaning of income data in the survey, but unfortunately—for the purposes of SAMOD—the cleaning is undertaken on net income rather than gross income. There is a gross income variable in the NIDS dataset, which is the primary source used for *yem* in SAMOD. However, during the data preparation stage for SAMOD it was evident that there were some large outliers when gross income was compared with cleaned net income. The SAMOD data preparation team undertook some work to address these outliers (which occurred throughout the income distribution) but were unable to reverse-engineer gross income from net income. Figure 2 shows that the under-reporting of gross income in the SAMOD underpinning dataset has a more-or-less linear relationship with earnings reported in the administrative tax data. This means that it may be possible to improve the earned income component of the SAMOD underpinning dataset; this is picked up in the conclusion and a way forward recommended.

As regards taxable income, Figure 3 shows the mean and median taxable income by taxable income decile for SAMOD and PITMOD. For the purposes of this figure, taxable income is defined as income after certain allowable deductions (for example in respect of allowable deductions relating to private pension contributions and interest).

Figure 3: Mean and median taxable income in 2017/18 by decile SAMOD and PITMOD



Note: the SAMOD and PITMOD cases include only those with a taxable income greater than zero.

Source: authors' calculations using SAMOD Version 7.4 (using NIDS Wave 5 Version 1.1) and PITMOD's administrative dataset.

As with earned income, taxable income is consistently underestimated in SAMOD. This is unsurprising given the findings relating to *yem* discussed above.

In addition to the general lack of coincidence between earned income/taxable income in SAMOD and PITMOD, there is also a high probability of both 'unit non-response' and 'item non-response' (Lavrakas 2008) in SAMOD, especially in relation to the 10th decile. It is known that, for example, as regards NIDS wave 5, there is a degree of unit non-response generated in part by attrition, particularly of wealthy individuals (Finn and Leibbrandt 2016; Rasmussen 2017).

4.2 Comparing simulations of PIT using SAMOD and PITMOD with published administrative aggregates

In this sub-section the simulated PIT from both PITMOD and SAMOD are compared with each other and set against aggregate-level published administrative data sources.

The interpretation of the tables in this section requires care, as simulated figures for both SAMOD and PITMOD do not precisely correspond to concepts used in the published administrative data. It is important to note that the simulated PIT in both SAMOD and PITMOD represent tax liability for the year in question, given the incomes of individuals for that year. In accountancy terms, this is referred to as being on an 'accrual' basis. On the other hand, the aggregate administrative data used for validation are published on a 'cash-flow' basis and so none of the reported figures precisely corresponds to what is simulated by SAMOD and PITMOD.

The most recent published statistics are derived from *Tax Statistics 2020*, published jointly by NT and SARS in December 2020. In terms of the number of individuals expected to submit returns for 2018, SARS estimates this to be 6,594,651. The SARS 'number of taxpayers assessed' figure (5,372,210) comprises only those taxpayers who have been assessed. This amounts to a percentage assessed of 81.5 (NT and SARS 2020: 37). The percentage of taxpayers assessed will increase over time. However, both the number of individuals expected to submit returns and the number of individuals assessed exclude those taxpayers that are not required to submit returns, who simply pay tax through PAYE. This includes the majority of taxpayers who have a gross employment income/salary below ZAR350,000 (the submission taxable income threshold).[13] To compound the issue, assessed individuals are not necessarily those who are due to pay tax in the year in question and may include those paying arrears of taxes or fines, or those receiving refunds. The figure will also not include those with tax liabilities for the current year that are not due to be paid until subsequent years. The reporting on a cash-flow basis will tend to overestimate revenue as it will include revenue due in previous years but not hitherto collected. It will also include fines and penalties. On the other hand, the fact that only a proportion of cases are assessed will result in an underestimate.

From Table 6 it can be seen that PAYE payments account for 96.4 per cent of tax collected in the tax year of interest—2017/2018 (shaded). This means that the large majority of PIT derives from employment income. To this are added provisional tax and assessment payments, accounting for 6.4 per cent and 3.5 per cent, respectively. Employment Tax Incentive (ETI) payments are deducted.[14] Tax refunds are also deducted but interest on overdue tax is added. As SAMOD and PITMOD do not model ETI, this should be added back when comparing the data with aggregate administrative data, but since interest on overdue tax is not tax per se, this should be deducted.

Table 6: Taxes on persons and individuals by year

	PAYE validating (R m)	Provisional Tax (R m)	Assessment payments (R m)	Employment Tax Incentive (ETI) (Rm)	Refunds (R m)	Subtotal (R m)	Interest on Overdue Tax (R m)	Total (R m)
2015/16	376,164	26,101	10,647	-4,063	-20,747	**388,102**	1,177	**389,280**
2016/17	410,807	28,641	12,719	-4,656	-22,965	**424,545**	1,379	**425,924**
2017/18	446,274	29,796	16,001	-4,317	-26,801	**460,953**	1,950	**462,903**
2018/19	477,503	34,935	14,668	-4,512	-30,511	**492,083**	1,746	**493,829**
2019/20	518,243	31,339	14,168	-4,754	-31,364	**527,633**	1,540	**529,172**
% of total								
2015/16	96.6%	6.7%	2.7%		-5.3%	**99.7%**	0.3%	**100.0%**
2016/17	96.5%	6.7%	3.0%		-5.4%	**99.7%**	0.3%	**100.0%**
2017/18	96.4%	6.4%	3.5%		-5.8%	**99.6%**	0.4%	**100.0%**
2018/19	96.7%	7.1%	3.0%		-6.2%	**99.6%**	0.4%	**100.0%**
2019/20	97.9%	5.9%	2.7%		-5.9%	**99.7%**	0.3%	**100.0%**

Source: NT and SARS (2020: 24, table A1.4.2: Taxes on persons and individuals, 20154/165—2019/20).

[13] Such individuals should also not have a car allowance/company car/travel allowance or other income (e.g. interest or rental income) and should not be claiming tax-related deductions/rebates (e.g. medical expenses, retirement annuity contributions other than pension contributions made by the employer, travel).

[14] ETI is a youth employment initiative that incentivizes employers to employ otherwise unemployed young people.

Drawing from Table 6, Table 7 compares the simulated PIT revenue for SAMOD and PITMOD with administrative data, having added back ETI and deducted interest on overdue tax. This table includes PAYE collection below the tax return threshold as well as assessed data.

Table 7: Reported and simulated revenue from personal income tax in 2017/18

	Reported (R m)	SAMOD		PITMOD	
		SAMOD simulated (R m)	% captured (simulated/reported)	PITMOD simulated (R m)	% captured (simulated/reported)
SARS	465,270	359,039	77.2	460, 439	99.0
National Treasury	460,953	359,039	77.9	460,439	99.9

Note: the SARS reported figures are derived from National Treasury and SARS (2019: 22, table A1.4.2); the NT reported figures are from the National Treasury Budget Report 2019.

Source: authors' calculations using SAMOD Version 7.4 and PITMOD Version 1.0 using 100% dataset.

From this table we can see that PITMOD performs well against reported tax revenue, despite the caveats mentioned earlier relating to the cash-flow versus accrual basis of the figures reported in published statistics. On the other hand, SAMOD captures only 77–78 per cent of PIT revenue. Interestingly, the data underpinning Figures 1 and 2 indicate that in all deciles mean gross earned income seems to be around 77 per cent lower than that derived from the administrative data.

4.3 Comparing simulations of taxpayers, taxable income, and PIT revenue using SAMOD and PITMOD

As has been indicated, comparisons with published aggregate administrative data are fraught with definitional challenges. Instead, in this section the output from SAMOD is compared with the output from PITMOD. This is closer to comparing like with like as both SAMOD and PITMOD simulate taxes on an accrual basis.

In Table 8 the output from each model is analysed according to taxable income groups.[15] Taxpayers are allocated to the taxable income group specified in the table according to their taxable income (including taxable income from lump sum payments). The table shows the number of taxpayers in each taxable income group as well as their total taxable income in ZAR million and the total PIT revenue in ZAR million. The values shown are annual amounts.

Overall, PITMOD and SAMOD yield similar totals for taxpayers and taxable income, with SAMOD generating 102 per cent of PITMOD's taxpayers, and 98 per cent of PITMOD's taxable income. However, SAMOD generates a much lower amount of PIT revenue: 79 per cent of that generated by PITMOD for the same period. This implies—in line with expectations—that PITMOD captures high-income individuals better than SAMOD, as well as simulating additional aspects of PIT that cannot be simulated in SAMOD. Breaking down the results by NT taxable income group, the greatest discrepancy between PITMOD and SAMOD occurs for the highest group (R1.5m and above): SAMOD has less than half the number of individuals in this top group (48 per cent of PITMOD), and simulates 55 per cent of PIT revenue compared with PITMOD.

[15] 'Taxable income group' is a categorization that is used by both NT and SARS for reporting purposes, though the groups they use vary. These taxable income groups do not correspond to the tax bands used for calculating PIT.

Table 8: Simulations of taxpayers and taxable income using SAMOD and PITMOD by NT taxable income group, 2018

NT taxable income	SAMOD			PITMOD			Ratio (SAMOD/PITMOD)		
group (R thousand)	Taxpayers (n)	Taxable income (R m)	Income tax (R m)	Taxpayers (n)	Taxable income (R m)	Income tax (R m)	Taxpayers	Taxable income	Income tax
R0–R70	26,822	363,819	107	157,160	197,421	5,755	0.2	1.8	0.0
R70–R150	2,609,601	365,136	14,042	1,984,350	267,628	13,084	1.3	1.4	1.1
R150–R250	1,823,559	360,803	36,014	1,758,410	355,340	37,113	1.0	1.0	1.0
R250–R350	915,991	272,132	40,935	1,131,060	335,667	51,342	0.8	0.8	0.8
R350–R500	678,947	282,270	55,112	851,160	351,891	69,680	0.8	0.8	0.8
R500–R750	456,380	280,021	69,910	517,290	312,246	79,012	0.9	0.9	0.9
R750–R1,000	223,773	192,329	56,294	204,670	175,548	52,498	1.1	1.1	1.1
R1,000–R1,500	85,189	99,253	32,057	130,520	155,820	52,052	0.7	0.6	0.6
R1,500 +	41,055	136,051	54,568	85,630	244,735	99,878	0.5	0.6	0.5
TOTAL	6,861,316	2,351,812	359,039	6,820,250	2,396,295	460,415	1.0	1.0	0.8

Note: amounts shown are annual figures; the data exclude income tax from lump sums.

Source: authors' calculations using SAMOD Version 7.4 and PITMOD Version 1.0 using 10% dataset.

Table 9: Simulations of taxpayers and taxable income using SAMOD and PITMOD by taxable income band, 2018

Taxable income band	**SAMOD**			**PITMOD**			**Ratio (SAMOD/PITMOD)**		
(R thousand)	**Taxpayers (n)**	**Taxable income (R m)**	**Income tax (R m)**	**Taxpayers (n)**	**Taxable income (R m)**	**Income tax (R m)**	**Taxpayers**	**Taxable income**	**Income tax**
0–189,880	3,468,148	872,969	26,452	2,918,120	602,464	30,938	1.2	1.4	0.9
189,881–296,540	1,427,634	335,741	40,023	1,571,450	379,245	48,059	0.9	0.9	0.8
296,541–410,460	796,027	271,522	45,766	982,340	341,156	59,025	0.8	0.8	0.8
410,461–555,600	455,633	212,178	44,871	578,540	273,741	59,499	0.8	0.8	0.8
555,601–708,310	325,514	203,744	51,488	293,790	183,360	47,313	1.1	1.1	1.1
708,311–1500,000	347,306	319,607	95,870	390,410	371,601	115,704	0.9	0.9	0.8
1500,001 +	41,055	136,051	54,568	85,630	244,735	99,878	0.5	0.6	0.5
TOTAL	6,861,316	2,351,812	359,039	6,820,280	2,396,302	460,416	1.0	1.0	0.8

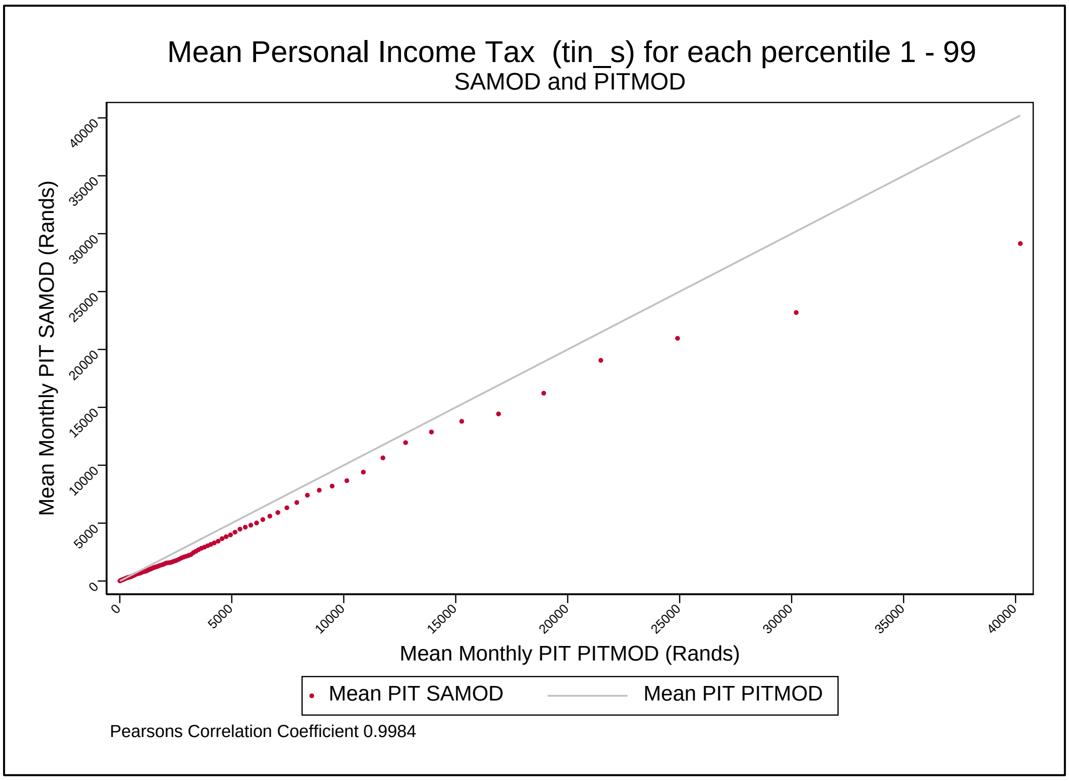
Note: amounts shown are annual figures; the data exclude income tax from lump sums; minor discrepancies in totals are due to rounding.

Source: authors' calculations using SAMOD Version 7.4 and PITMOD Version 1.0 using 10% dataset.

Table 10 makes the same comparison but by the tax bands that were applicable at the relevant timepoint. Again, a taxpayer is allocated to a band on the basis of their taxable income, meaning that although a taxpayer in a higher band will have paid tax in the lower bands, they appear only in the highest band applicable to their taxable income and all the tax payable by that taxpayer is allocated to that band. The findings are broadly similar to those in Table 9, with the highest tax band having the greatest discrepancy.

As was the case for earnings, it is possible to look at the relationship between PIT as simulated by SAMOD and PITMOD by calculating the mean monthly tax for each percentile of tax simulated using each of the two models. Figure 4 shows a scatter plot for mean income for percentiles 1–99 (percentile 100 is omitted because it is an extreme outlier).

Figure 4: Mean PIT (tin_s) for each percentile 1–99, SAMOD and PITMOD



Note: the SAMOD and PITMOD cases include only those with a PIT liability greater than zero and exclude cases with capital gains.

Source: authors' calculations using SAMOD Version 7.4 and PITMOD Version 1.0.

As can be seen from Figure 4, there is an almost linear relationship between the percentiles of mean monthly tax payable using SAMOD and the percentiles of mean monthly tax payable using PITMOD. Indeed the Pearson's correlation coefficient is 0.9984. This is similar to the picture that emerged when comparing mean monthly and income percentiles in Figure 2. Again, it suggests that the gross earned income in the data underpinning SAMOD is under-counted across the distribution. As previously indicated, a possible mechanism to address this is suggested in the conclusion.

4.4 Comparing simulations of PIT revenue between PITMOD and administrative data

In this final section, the performance of PITMOD is validated against information contained in the source administrative data. The source administrative data contain information on assessed income tax for most of the assessed cases. For the non-assessed cases, tax collected through the PAYE system is reported.

In order to check the performance of PITMOD, a new variable was created in the administrative data containing the assessed tax. Where this was missing or zero, the reported PAYE amount was substituted.

Flags for two anomalous scenarios were created. First, where cases should have been assessed because the employed earners had a gross salary that exceeded ZAR350,000 but there was no assessed information in the data, a flag was set indicating that these cases would likely be assessed in the future: 2.3 per cent of cases were in this category. Of course, other cases could have been pending assessment in addition to those above the salary limit, but as the rules governing assessment are complex, it was not possible to identify these cases definitively. Second, cases were identified with anomalous age-related tax rebates and/or medical tax credits that were inconsistent with the rules in place at that time. Accordingly, a further flag was set to signal these cases (11.87 per cent of all cases).

These flags were used when benchmarking the performance of PITMOD against the source administrative data. In Table 10, the monthly income tax calculated by PITMOD is compared with that recorded in the administrative data. The percentage of cases where the tax estimated by PITMOD fell within 10 Rand, 100 Rand, and 200 Rand were calculated. The results are presented in the first instance with no cases excluded from the administrative data. Next, the results are presented excluding cases from the administrative data where, on an earnings basis, assessment should have taken place but had not yet occurred. Finally, the results are presented excluding both the cases that should have been assessed and cases where there were inconsistent tax rebates.

Table 10: A comparison of PITMOD's simulated personal income tax and that recorded in the administrative data

	Within ZAR10 per month (%)	Within ZAR100 per month (%)	Within ZAR200 per month (%)
All cases (no cases excluded)	78.27	85.61	89.37
Excluding PAYE cases over ZAR350,000 pa not assessed (2.3% of cases)	79.87	86.98	90.52
Excluding PAYE cases over ZAR350,000 pa not assessed (2.3% of cases) and those with irregular rebates (an additional 11.86% of cases)	87.35	92.08	94.37

Source: authors' construction based on PITMOD and raw administrative data.

Even where no cases have been excluded, PITMOD performs reasonably well, with nearly 90 per cent of cases falling within ZAR200 per month of the figures recorded in the administrative data. Taking into account exclusions, performance was significantly enhanced. These results are encouraging bearing in mind that foreign tax rebates could not be modelled within PITMOD, and also that PITMOD does not fully take into account tax adjustments from previous years.

5 Concluding remarks

This paper has presented results from two microsimulation models about PIT. One model is underpinned by a nationally representative household survey (SAMOD) and the other by anonymized tax records (PITMOD). PITMOD enables the PIT system to be simulated much more precisely as it can be calculated in fine detail. It was important to explore the extent to which

PITMOD's results correspond to published administrative totals about PIT for the same period as this needs to be ascertained before policy reforms can be simulated. It was also important to examine the extent to which the simulated PIT results compare with those generated by SAMOD using a household survey as this should inform the extent to which the direct taxes simulated in SAMOD can be interpreted.

At an aggregate level, it was found that PITMOD simulates between 99 and 100 per cent of reported PIT revenue for the relevant period (the 2018 tax year) (Table 7). At an individual case level, PITMOD's simulated values for PIT correspond well with SARS' raw tax liability data: when comparing all cases, 78 per cent matched within ZAR10 per month, rising to 89 per cent matching within ZAR200 per month. If cases are excluded that had not yet been assessed, or that had anomalous rebates in the source administrative data, 87 per cent of cases matched within ZAR10 per month, rising to 94 per cent matching within ZAR200 per month (Table 10).

In terms of PITMOD's accuracy, therefore, it can be concluded that it simulates the tax legislation and calculation steps very well. This means that analysis of the functioning of different aspects of the PIT system can be explored using PITMOD with a large degree of confidence. In addition, policy reforms of the PIT system would best be undertaken using PITMOD. In practice a 10 per cent sample of the PITMOD input data will be sufficient for most analyses.

On the other hand, SAMOD captures only 77–78 per cent of reported PIT revenue for the relevant period (Table 7). When comparing PITMOD and SAMOD (Tables 8 and 9), the two models yield similar totals for taxpayers and taxable income, with SAMOD generating 102 per cent of PITMOD's taxpayers, and 98 per cent of PITMOD's taxable income. However, SAMOD generates a much lower amount of PIT revenue: 79 per cent of that generated by PITMOD for the same period. Breaking down the results by NT taxable income group, the greatest discrepancy between PITMOD and SAMOD occurs for the highest group (R1.5m and above): SAMOD has less than half the number of individuals in this top group (48 per cent of PITMOD), and simulates only 55 per cent of PIT revenue compared with PITMOD.

In terms of SAMOD's accuracy, it is evident that it captures only around three-quarters of actual PIT revenue. There are ways in which this could be addressed. As has been indicated, the NIDS team undertake extensive cleaning of the net earned income reported but not of the gross income variable. Although the gross income variable was cleaned by the SAMOD team, there are other mechanisms that can be brought to bear now that there is a more detailed profile of the discrepancies. For example, it would be possible to impute gross incomes from net incomes on a case-by-case basis using an iterative process involving the repeated calling of SAMOD from STATA. Alternatively, a model could be fitted based on the characteristics of the groups having the discrepancies, to adjust their gross incomes.

The PITMOD model will be used in-house by SARS, and in-house training sessions will need to be arranged, supplemented by a detailed manual and an automated summary statistics module. It is intended that the model will also be made available more generally for use by government and researchers, possibly via the NT data secure room. Users would also need training on how to use and interpret the model. PITMOD will provide an important resource for research and policy.

Acronyms

ISER	Institute for Social and Economic Research, University of Essex
ETI	Employment Tax Incentive
LCS	Living Conditions Survey
NIDS	National Income Dynamics Study
NT	National Treasury
PAYE	Pay-As-You-Earn
PIT	Personal Income Tax
PITMOD	A microsimulation model of South Africa's Personal Income Tax policy, underpinned by administrative data
SAMOD	A South African tax–benefit microsimulation model, underpinned by survey data
SARS	South African Revenue Service
SASPRI	Southern African Social Policy Research Insights
UIF	Unemployment Insurance Fund
UNU-WIDER	United Nations University World Institute for Development Economics Research
VAT	Value-added Tax

References

Arndt, C. (2018). 'New Data, New Approaches and New Evidence: a Policy Synthesis'. *South African Journal of Economics*, 86(1): 167–78. https://doi.org/10.1111/saje.12184

Bassier, I., and I. Woolard (2020). 'Exclusive Growth?: Rapidly Increasing Top Incomes amidst Low National Growth in South Africa'. WIDER Working Paper 2020/53. Helsinki: UNU-WIDER.

Bover, O., J.M. Casado, E. Garcia-Miralles, J.M. Labeaga, and R. Ramos (2017). 'Microsimulation Tools for the Evaluation of Fiscal Policy Reforms at the Banco de España'. Documentos Ocasionales 1707. Madrid: Banco De España. https://doi.org/10.2139/ssrn.3048727

Branson, N., and M. Wittenberg (2019). 'Longitudinal and Cross-Section Weights in the NIDS Data 1–5'. Technical Paper 9. Cape Town: National Income Dynamics Study.

Ceriani, L., C.V. Fiorio, and C. Gigliarano (2013). 'The Importance of Choosing the Data Set for Tax–Benefit Analysis'. *International Journal of Microsimulation*, 6(1): 86–121. https://doi.org/10.34196/ijm.00078

Ebrahim, A., and C. Axelson (2019). 'The Creation of an Individual-Level Panel Using Administrative Tax Microdata in South Africa'. SA-TIED Working Paper 36. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/661-6

Figari, F., A. Paulus, and H. Sutherland (2014). 'Microsimulation and Policy Analysis'. ISER Working Paper 2014-23. Colchester: Institute for Social and Economic Research, University of Essex.

Finn, A., and M. Leibbrandt (2016). 'The Dynamics of Poverty in the First Four Waves of NIDS'. SALDRU Working Paper 174/NIDS Discussion Paper 2016/1. Cape Town: SALDRU, University of Cape Town.

Flory, J., and S. Stöwhase (2012). 'MIKMOD-ESt: a Static Microsimulation Model of Personal Income Taxation in Germany'. *International Journal of Microsimulation*, 5(2): 66–73. https://doi.org/10.34196/ijm.00073

Hundenborn, J., J. Jellema, and I. Woolard (2017). 'Income Inequality and Taxation: the Case of South Africa'. Poster presented at WIDER Development Conference, Maputo.

Hundenborn, J., I. Woolard, and J. Jellma (2019). 'The Effect of Top Incomes on Inequality in South Africa'. *International Tax and Public Finance*, 26: 1018–47. https://doi.org/10.1007/s10797-018-9529-9

Jelloul, M.B., A. Bozio, T. Douenne, B. Fabre, and C. Leroy (2019). 'Le modèle de micro-simulation TAXIPP - Version 1.1'. Guide méthodologique. Paris: Institut des Politiques Publiques.

Lavrakas, P.J. (ed.) (2008). *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781412963947.n298

Leventi, C., M. Matsaganis, and M. Flevotomou (2013). 'Distributional Implications of Tax Evasion and the Crisis in Greece'. EUROMOD Working Paper EM17/13. Colchester: University of Essex.

McLennan, D., M. Noble, M. Mpike, G. Wright, and C. Byaruhanga (2017). 'South Africa Microdata Scoping Study 2016. Report for the Monitoring and Learning Facility of the Programme to Support Pro-Poor Policy Development'. Presidency, Republic of South Africa and Delegation of the European Union.

Ministry of Finance (2013). 'Minister Gordhan Announces Further Details of the Tax Review Committee and the Terms of Reference'. Media Statement, 17 July. Pretoria: Ministry of Finance, Republic of South Africa. Available at: http://www.treasury.gov.za/TaxReviewCommittee/ (accessed 27 April 2021).

Miola, E., and M. Manzo (2021). 'A Tax–Benefit Microsimulation Model for Personal Income Taxation in Italy'. DF WP 10. Rome: MEF Ministero dell-Economia e delle Finanze.

Mirrlees, J., S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles, and J.M. Poterba (2011). *Tax by Design. The Mirrlees Review*. Oxford: Oxford University Press.

National Treasury (2014). 'Budget Review 2014'. Government Printer [Online]. Available at: http://www.treasury.gov.za (accessed 24 April 2021).

National Treasury (2015). 'Budget Review 2015'. Pretoria: National Treasury South Africa.

National Treasury and SARS (2019). 'Tax Statistics 2019'. Pretoria: National Treasury and South African Revenue Service.

National Treasury and SARS (2020). 'Tax Statistics 2020'. Pretoria: National Treasury and South African Revenue Service.

OECD (2013). 'Policy Framework for Investment User's Toolkit', Chapter 5: Tax Policy. Paris: Investment Division of OECD Directorate for Financial and Enterprise Affairs.

Orthofer, A. (2016). 'Wealth Inequality in South Africa: Evidence from Survey and Tax Data'. REDI3x3 Working Paper 15. Research Project on Employment, Income Distribution and Inclusive Growth. Cape Town: SALDRU, University of Cape Town.

Rasmussen, E.H. (2017). 'Increasing Progressivity in South Africa's Personal Income Tax System'. Thesis for the degree of Master of Commerce. Cape Town: University of Cape Town.

SALDRU (2014). 'National Income Dynamics Study 2014–2015'. Wave 4 [dataset]. Version 1.1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016. Pretoria: Department of Planning Monitoring and Evaluation [commissioner], 2014.

SALDRU (2018). 'National Income Dynamics Study 2017'. Wave 5 [dataset]. Version 1. Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town: Southern Africa Labour and Development Research Unit [implementer]. Cape Town: DataFirst [distributor]. https://doi.org/10.25828/fw3h-v708.

Shine, M., C. Clark, R. Tonkin, and D. Webber (2019). 'Using Tax Data to Better Capture Top Earners in Household Income Inequality Statistics'. London: Office for National Statistics.

Stats SA (2017a). 'Living Conditions of Households in South Africa: An analysis of household expenditure and income data using the LCS 2014/2015'. Statistical Release P0310 (2015). Pretoria: Statistics South Africa.

Stats SA (2017b). 'Inequality Trends in South Africa: a Multidimensional Diagnostic of Inequality'. Report 03-10-19. Pretoria: Statistics South Africa.

Stats SA (2017c). 'Consumer Price Index August 2017'. Statistical Release P0141. Pretoria: Statistics South Africa.

Steenekamp, T.J. (2012). 'The Progressivity of Personal Income Tax in South Africa Since 1994 and Directives for Tax Reform'. *South African Business Review*, 16(1): online. Available at: https://www.ajol.info/index.php/sabr/article/view/85455 (accessed 3 June 2021).

Sutherland, H. (2001). 'EUROMOD: an Integrated European Benefit–Tax model – Final Report'. EUROMOD Working Paper EM9/01. Colchester: University of Essex.

Sutherland, H. (2018). 'Quality Assessment of Microsimulation Models: the Case of EUROMOD'. *International Journal of Microsimulation*, 11(1): 198–223. https://doi.org/10.34196/ijm.00178

Sutherland, H., and F. Figari (2013). 'EUROMOD: the European Union Tax–Benefit Microsimulation Model'. *International Journal of Microsimulation*, 6(1): 4–26. https://doi.org/10.34196/ijm.00075

Tammik, M. (2018). 'Baseline Results from the EU28 EUROMOD (2014-2017)'. EUROMOD Working Paper EM5/18. Colchester: University of Essex.

Tanzi, V. (2001). *Tax Policy for Developing Countries*. Washington, DC: IMF.

Tanzi, V. (2004). 'Globalization and the Need for Fiscal Reform in Developing Countries'. Special Initiative on Trade and Integration. Inter-American Development Bank.

University of Essex (2019). 'EUROMOD software v3.1.8'. Colchester: University of Essex.

Verbist, G., and N. Van Mechelen (2020). 'BELMOD: Adapting EUROMOD for the Use of Administrative Data in Belgium'. Presentation for Annual Meeting. Antwerp: Universiteit Antwerpen.

Wilkinson, K. (2009). 'Adapting EUROMOD for Use in a Developing Country – the Case of South Africa and SAMOD'. EUROMOD Working Paper EM5/09. Colchester: University of Essex.

Wittenberg, M. (2017). 'Measurement of Earnings: Comparing South African Tax and Survey Data'. SALDRU Working Paper 212. Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town.

Wright, G., M. Noble, M. Dinbabo, P. Ntshongwana, K. Wilkinson, and P. Le Roux (2011). 'Using the National Income Dynamics Study as the Base Micro-dataset for a Tax and Transfer South African Microsimulation Model'. Report produced for the Office of the Presidency, South Africa. Oxford: Centre for the Analysis of South African Social Policy, University of Oxford.

Wright, G., H. Barnes, M. Noble, D. McLennan, and F. Masekesa (2018). 'Assessing the Quality of the Income Data used in SAMOD: a South African Tax–Benefit Microsimulation Model'. WIDER Working Paper 2018/173. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2018/615-9

Appendix A: Variables used in SAMOD for the simulation of PIT[16]

Table A1 summarizes the different income or expenditure-related variables that are used in the PIT policy in SAMOD, and specifies whether they are derived from the NIDS survey data using self-reported information ('Survey'), or are generated on-model ('SAMOD').

Table A1: Types of income or expenditure taken into account or generated in the PIT simulations in SAMOD

Type	Obtained from self-reported survey data or simulated on-model?	Variable name
Income from employment	Survey	yem
Income from self-employment	Survey	yse
Income from property	Survey	ypr
Income from private pension	Survey	ypp
Income (other)	Survey	yot
Expenditure on health insurance by employer	Survey	xishler
Employee's contribution to UIF	SAMOD	tscee_s
Income from interest payments	Survey	yiyit
Income from interest payments above threshold	SAMOD	ttaiy_s
Expenditure on pension contributions	Survey	xpp
Tax deduction for pension contributions	SAMOD	ttapn_s
Income from retirement-related lump sum	Survey	yivls
Income from retirement-related lump sum above threshold	SAMOD	ttaoy_s
Income from employment-related lump sum	Survey	ysv
Income from employment-related lump sum above threshold	SAMOD	ttasv_s
Tax payable on retirement or employment-related lump sum	SAMOD	tinkt_s
Expenditure on medical expenses (not insurance)	Survey	xhl
Expenditure on medical scheme	Survey	xishl
Medical scheme fees tax credit	SAMOD	tintchl_s
PIT rebate	SAMOD	tinta_s
General taxable income	SAMOD	ttb_s
PIT payable	SAMOD	tin_s

Source: authors' compilation using SAMOD Version 7.3.

The PIT calculation in SAMOD can be summarized as follows:

Income tax payable = Tax payable on (general taxable income + Taxable income from interest payments – Tax deductions for pension contributions) + Tax payable on lump sums – Tax rebate – Medical tax credits

Using the nomenclature of the SAMOD variables, the amount of PIT payable, tin_s, can therefore be summarized as follows:

tin_s = Tax payable on (il_taxabley[17] *+ ttaiy_s - ttapn_s)*[18] *+ tinkt_s – tinta_s – tintchl_s*

[16] This is a modified version of appendix 1 in Wright et al. (2018), updated where applicable to reflect the timepoint and model version (SAMOD V7.4) used in this paper.

[17] *Il_taxabley* is calculated as the sum of *yem, yse, ypr, ypp, yot,* and *xishler.*

[18] In combination this equals the composite variable *ttb_s.*

The variable *tinkt_s* is the tax on retirement-related and employment-related lump sums. Taking into account variants to the rules for those aged less than 55, these taxes are simulated on-model within SAMOD for reported lump sums in excess of the threshold of ZAR500,000 per year (in variables *ttaoy_s* for retirement-related lump sums and *ttasv_s* for employment-related lump sums). The excess lump sum amounts (*ttaoy_s* + *ttasv_s*) are taxed at 18 per cent for amounts less than or equal to ZAR200,000 per year; 27 per cent for amounts of ZAR200,001 to ZAR550,000 per year; and 36 per cent for amounts of ZAR550,001 and over per year.

The general tax rebate (*tinta_s*) is simulated on-model within SAMOD for 2017 at ZAR13,635 per year, with an additional rebate of ZAR7,479 for those aged 65 and over, and a further ZAR2,493 for those aged 75 and above. These tax rebates are deductions from tax calculated rather than tax-free thresholds.

The medical tax credits (*tintchl_*s) are simulated on-model and comprise two elements: the medical scheme fees tax credit, which is calculated for the medical scheme contributor (R270 per month), their first dependant (R270 per month), and any additional dependants (R181 per month each); and the additional medical expenses tax credit (excess medical scheme fees and qualifying medical expenses), taking into account the variants to the rules for those who were aged 65 or over and/or had a spouse or child with a disability. Again, the total amount of medical tax credit is deducted from the amount of tax payable.

The final amount of tax payable (*tin*_s) is calculated in SAMOD for June 2017 at 18 per cent for the first ZAR189,880 per year; 26 per cent for ZAR189,881 to ZAR296,540; 31 per cent for ZAR296,541 to ZAR410,460; 36 per cent for ZAR410,460 to ZAR555,600; 39 per cent for ZAR555,601 to ZAR708,310; 41 per cent for ZAR708,311 to ZAR1,500,000; and 45 per cent for amounts above ZAR1,500,000.

Appendix B: Constructing PITMOD's input dataset

This appendix provides further information about how the source administrative data were harnessed to create the variables needed to construct PITMOD's input dataset.

B1 Income tax formsf and tax assessment process

SARS calculates an individual's personal income tax liability for a given tax year on the basis of one or more tax-related forms submitted by the individual and/or their employer, with the number and type(s) of form(s) dependent upon the individual's personal employment circumstances and other relevant financial circumstances. The tax forms that are of greatest importance to this project are the 'IRP5/IT3a employee tax certificate' and the 'ITR12 personal income tax return'.

IRP5/IT3a: Employee tax certificates submitted to SARS by the employer

For people engaged in paid employment for an employer, the starting point is typically the submission of an IRP5 or IT3a employee tax certificate. These forms are submitted by the employer on behalf of the employee. The IRP5 form is submitted for employees who have tax deducted through the pay-as-you-earn (PAYE) system while the IT3a form is submitted for employees who receive a wage/salary but for whom no tax is deducted through the PAYE system. These forms record any remuneration received by the employee and/or any lump sums received by the employee from the employer, pension fund, provident fund, or retirement annuity fund. For those individuals who are employed by more than one employer during the course of a given tax year, either concurrently or at different times, each employer is required to submit a separate IRP5/IT3a form.

ITR12: Personal income tax return

The ITR12 form is completed by individuals with employee income over ZAR350,000, individuals who work for more than one employer during the given tax year, individuals who have additional sources of income or tax-related deductions or rebates that are not taken into account in their IRP5/IT3a employer return, and individuals who are not employees (and therefore do not have an IRP5/IT3a form) and are not part of the PAYE system.[19] Income from sources such as self-employment business activities, property rental income, investment income (e.g. interest and dividends), and capital gains must be recorded on an ITR12 personal income tax return form. When an individual starts to complete an ITR12 form, they will see that certain fields have been automatically pre-populated with values from any relevant IRP5/IT3a forms that relate to them. The ITR12 form therefore consolidates any information from the IRP5/IT3a forms and provides an opportunity for the individual to enter details of additional expenses and/or income from sources not covered on the IRP5/IT3a forms.

[19] Individuals are not required to complete an ITR12 form if they satisfy *all* of the following criteria:

- Their total employment income for the tax year before tax (i.e. gross income) was ZAR350,000 or less.
- They received employment income from only a single employer during the tax year.
- They did not have any other forms of income (e.g. business or investment income) and did not receive a car allowance or company car or travel allowance.
- They did not claim any tax-related deductions or tax rebates (e.g. medical expenses, retirement annuity contributions other than pension contributions made direct by their employer).

.

Some of the information required for the ITR12 form is already contained within other tax forms that the individual must complete. For instance, additional information on investment income is contained within form IT3b, and so this information is pre-populated on the ITR12 form. Similarly, any information received from third-party organizations, primarily related to medical insurance scheme contributions, is automatically pre-populated on the ITR12 form.

People with relatively straightforward income, employment and tax affairs may not need to complete an ITR12 form as all the information needed by SARS to compute tax liability is contained within a single IRP5/IT3a form.

Tax assessment

Tax assessment is the process through which SARS calculates the final tax liability for an individual. Any individuals who are not required to submit an ITR12 form are therefore not required to undergo the tax assessment process as all their tax-related information is contained within the IRP5/IT3a forms and, if they are liable to pay any tax on their employment income, this amount is automatically deducted at source through the PAYE system. In contrast, those individuals who are required to submit an ITR12 form undergo tax assessment in order to ensure that all relevant incomes and deductions are properly taken into account. The submission of the ITR12 personal income tax return therefore forms the basis of the tax assessment process.

During the tax assessment process, SARS may revise (upwards or downwards) the values of individuals' incomes, allowances, deductions, etc. and/or they may reclassify values from one category to another depending on the totality of the information submitted through the ITR12 personal income tax return. The final configuration of an individual's income tax affairs can therefore be significantly different from the data submitted in the income tax returns, as errors and omissions are corrected during the tax assessment process.

B2 Constructing the source administrative data file

The source administrative data file was built by the Data Team at SARS especially for the purpose as there was not a readily available 'off-the-shelf' dataset already in existence that met the requirements of the PITMOD project.

The source administrative data file was constructed by bringing together various existing SARS data sources and supplementing these with third-party information on medical insurance scheme contributions. The SARS data were drawn from two systems, each comprising a number of datasets:

- IRP5/IT3a data;
- ITR12 data.

The third-party medical insurance scheme data, which are used by SARS in the assessment process, were added to the above data sources.

Table B1 summarizes the data sources for the source administrative data file.

Table B1: Summary of data sources for the source administrative data file

IRP5	IRP5IT3A	This table contains the details of the IRP5/IT3A employee tax certificates as issued
	IRP5IT3AAMOUNTDETAIL	This table, associated with the table IRP5IT3A, contains details of income received, deductions made, contributions made, tax withheld, etc., as taken from the employee tax certificates (IRP5/IT3A)
ITR12	IT Return	This table contains details of the income tax returns as submitted by taxpayers
	IT-RETURN-AMT	This table contains details of assets, losses, debits, and income amounts, etc., as submitted by taxpayers on the tax return
	ASSESSMENT	This table contains income tax assessment details as supplied on the tax return
	ASSESSED-AMT	This table contains the amount details required for the assessment of the taxpayer

Source: authors' construction.

One specification for the source administrative data file was that it must be configured such that each individual case in the dataset relates to a single individual. As people can legitimately have more than one IRP5/IT3a submission in any given tax year, it was therefore necessary for SARS to calculate a single overall composite IRP5/IT3a return for each individual who had multiple IRP5/IT3a forms. This was achieved by creating a unique identifier for individuals in the IRP5/IT3a data from identifiers found in the dataset using the following variables in the order of completeness: tax reference number, ID number, passport number, and certificate number.

The next step entailed SARS taking the compiled IRP5/IT3a dataset produced in the first step and merging with it information from the ITR12 returned data based on the unique Tax Reference Number and/or unique ID number. This step added information on additional incomes, allowances, deductions, etc. As would be expected, not all cases in the IRP5/IT3a data were matched to information in the ITR12 returned data because not all employees are required to submit a personal income tax return. Similarly, there were some cases in the ITR12 data that did not match the IRP5/IT3a dataset, which is also to be expected as some people are obliged to complete a personal income tax return despite not having a submitted IRP5/IT3a form (e.g. people with business income or investment income who are not employees). The composition of the data file at the end of this step therefore consists of people who had a returned IRP5/IT3a submission but no ITR12, people who had a returned ITR12 submission but no IRP5/IT3a, and people who had both returned IRP5/IT3a and returned ITR12 submissions. Table B2 shows the numbers of cases in each of these three groups.

Table B2: Number of cases in the dataset following the merging of IRP5/IT3a and ITR12 data

	Number of cases	% of total cases
Only IRP5/IT3a returned data	9,583,689	65.06
Only ITR12 returned data	369,810	2.51
Both IRP5/IT3a and ITR12 returned data	4,777,268	32.43
Total cases in the dataset	14,730,767	100

Source: PITMOD (after deleting 32 duplicate cases and 60 implausible cases).

As part of the assessment process carried out on the ITR12 returned data, third-party medical scheme data are utilized as these feed into the calculation of medical tax credit rebates. Medical scheme data are held in a separate data system within SARS and so they were merged into the source administrative data file using the Tax Reference Number and/or unique ID number. The information provided by the third-party medical insurance scheme organizations consists of the number of scheme members and dependants registered in the scheme(s) at the beginning of the tax year and any changes to this number of registered individuals throughout the course of the tax year. Not all individuals who are required to submit an ITR12 personal income tax return subscribe to medial insurance cover, but subscribing to a medical insurance scheme would in itself require the person to complete an ITR12 form.

The fourth and final step involved adding information to the source administrative data file concerning the outcome of the tax assessment process for those individuals who have been through the process. As noted above, the values presented in the assessed data may be the same as those submitted in the returned data, or they may differ due to revisions having been made during the assessment process. Table B3 shows the number and proportion of cases in each of the three groups (IRP5/IT3a only; ITR12 only; IRP5/IT3a and ITR12) that have been through the tax assessment process and therefore for whom information is available from the assessed data.

Table B3: Number of cases having completed the tax assessment process

	Number of cases	Number of cases having been through the tax assessment process	% of cases having been through the tax assessment process
Only IRP5/IT3a returned data	9,583,689	n/a	n/a
Only ITR12 returned data	369,810	369,810	100%
Both IRP5/IT3a and ITR12 returned data	4,777,268	4,461,967	93.4%
Total cases in the source administrative data file	14,730,767	4,831,777	32.8%

Source: authors' construction.

A combination of assessed and returned data is used in the derivation of variables for the microsimulation model, as detailed in Table B4 at the end of this appendix.

The construction of the source administrative data file was an iterative process involving the drawing of small anonymized samples that were explored in detail by the research team, which led to incremental improvements to the specification of the data file. The key objective of this iterative process was to ensure that the source administrative data file contained all the necessary variables from the returned data and the assessed data to enable individuals' final tax liability to be accurately calculated.

B3 Variables in the source administrative data file

The final source administrative data file produced by SARS from which to construct a dataset to underpin the PITMOD microsimulation model consists of a mixture of variables drawn from the IRP5/IT3a and ITR12 returned data, third-party medical insurance scheme data, and ITR12 assessed data. In total there were 1,390 variables in the source administrative data file.

Some variables relate to particular 'source codes' as specified in the SARS tax guidance documentation; for example, source code 3601 relates to taxable wages/salary. Other variables are derived by SARS through the combination of different source codes to generate meaningful summary variables; for example, variables with the prefix business_inc_p are the aggregate of source codes relating to local business income (profit). In addition to the monetary values contained within the returned and assessed data there are a few demographic variables sourced from the ITR12 returned data, while the medical insurance scheme data reflect the number of members and dependants registered for the scheme(s).

The variables that relate to tax system source codes (e.g. 'code3601', which relates to taxable wages/salaries) follow a relatively straightforward naming convention. All source code variables have a four-digit numeric nomenclature, with the first two digits indicating the broad category of tax-related information, and the following two digits indicating the detailed sub-category of tax-related information. The same source codes are used in the returned data and the assessed data but the variables have either the suffix '_return' or the suffix '_assd' to identify whether they are returned or assessed values.

B4 Variables in PITMOD's input dataset

Table B4 lists each of the variables in PITMOD's input dataset, including the source code or variable from which it was derived, the variable name in PITMOD, and whether or not the variable is transformed on-model in some way before being either added to taxable income (see 'Income' section of the table), deducted from taxable income (see 'Deductions' section of the table), or subtracted from tax liability (see 'Tax credits' section of the table).

In addition, an age variable (*dag*) was constructed from the variable *age*. Age was missing for 0.43 per cent of cases in the full dataset and had to be imputed using variables relating to rebates (*rebagecode*, *amtrebageval*, and *amtrebatetertiaryvalue*). Missing ages were imputed as 30, 70, and 80 depending on whether information for primary, secondary, or tertiary rebates was present in the data.

A married in community of property flag (*dms*) was created from the variables *marriedind* and *marriagetype*. This variable is coded 0.5 where an individual is married in community of property and 1 otherwise.

Table B4: Variables in PITMOD's input dataset

Element of PIT	Source code(s)/variable(s)	PITMOD variable name	Transformed in PITMOD? (Policy name and output variable)
Income			
Employment income			
Employee income	For all variables the assessed data are used; if no assessed data, then returned data (ITR12 or IRP5) are used		Y (inc_employee) yem_s
Salaries and wages (may include incentive awards)	3601 (local) 3651 (foreign)	yemwg yemabwg	
Allowances	370x, 371x, 372x (local) excluding non-taxable and individual codes identified below	yemal	
	375x, 376x, 377x (foreign) excluding non-taxable and individual codes identified below	yemabal	
			Expenses deducted from allowances:
	3701, 3702 (local)	yemaltr	yemaltr_s
	3704 (local)	yemalsu	yemalsu_s
	3715, 3754, 3765 (foreign)	yemabalsu	yemabalsu_s
	3708 (local)	yemalpo	yemalpo_s
Fringe benefits	380x, 381x, 382x, 383x (local) excluding non-taxable	yemfb	
	385x, 386x, 387x, 388x (foreign) excluding non-taxable	yemabfb	

Element of PIT	Source code(s)/variable(s)	PITMOD variable name	Transformed in PITMOD? (Policy name and output variable)
Overtime	3607 (local) 3657 (foreign)	yemxp yemabxp	
Bonuses (may include incentive awards)	3605 (local) 3655 (foreign)	yemxb yemabxb	
Restraint of trade	3613 (local) 3663 (foreign)	yemrd yemabrd	
Arbitration award	3608 (local) 3658 (foreign)	yemaw yemabaw	
Independent contractor	3616 (local) 3666 (foreign)	yemic yemabic	
Labour broker	3617, 3619 (local) 3667, 3669 (foreign)	yemlb yemablb	
Pensions	3603 (local) 3653 (foreign)	ypp yppab	
Annuities	3610, 3611 (local) 3660, 3661 (foreign)	ypa ypaab	
Director fees	3615, 3620 (local) 3665, 3670 (foreign)	ydf yabdf	
Commission	3606 (local) 3656 (foreign)	yco yabco	Expenses deducted from commission income: yco_s
Business income	The assessed data are used		
Local business profit/loss	business_inc_p_assd business_inc_p_lrfcya business_inc_l_lrfcya	yse	Y (inc_business) yse_s
Foreign business profit/loss	4222	yseab	Y (inc_business) yse_s
Farming income	The assessed data are used		
Local farming profit/loss	agri_businc_p_loc_assd agri_businc_p_loc_lrfcya agri_businc_l_loc_lrfcya	yag	Y (inc_farming) yag_s
Foreign farming profit/loss	agri_businc_p_frgn_assd agri_businc_p_frgn_lrfcya agri_businc_l_frgn_lrfcya	yagab	Y (inc_farming) yag_s
Investment income	A combination of assessed and returned data is used		
Local interest[20]	4201	yiyit	Y (inc_interest) yiyit_s
Foreign interest[21]	4218	yiyabit	Y (inc_interest) yiyit_s
Local dividends—REIT and deemed dividends[22]	4238 4292	yiydv01 yiydv02	Y (inc_dividends) yiydv_s
Foreign dividends profit[23]	4216	yiyabdv (profit)	Y (inc_dividends) yiydv_s
Local capital gains/losses	4250 4251	ykg (profit) ykl (loss)	Y (inc_capital_gains) ykg_s
Foreign capital gains/losses	4252 4253	yabkg (profit) yabkl (loss)	Y (inc_capital_gains) ykg_s
Local rental income profit/loss	4210 4211	ypr (profit) yprlo (loss)	Y (inc_rental_income) ypr_s
Foreign rental income profit/loss	4288 4289	yprab (profit) yprablo (loss)	Y (inc_rental_income) yprab_s

[20] The variables *dag* and *dms* are also used.

[21] The variable *dms* is also used.

[22] The variable *dms* is also used.

[23] The variable *dms* is also used.

Element of PIT	**Source code(s)/variable(s)**	**PITMOD variable name**	**Transformed in PITMOD? (Policy name and output variable)**
Other taxable income	For all variables the assessed data are used; if no assessed data then returned data (ITR12) are used		
Royalties profit/loss	4212 (local) 4213 (local) 4278 (foreign) 4279 (foreign)	yro (profit) yrolo (loss) yabro (profit) yabrolo (loss)	N
Other	4214 (local) 4220, 4228 (foreign)	yot yotab	N
Deductions	Except where otherwise stated, the assessed data are used; if no assessed data then returned data (ITR12 or IRP5) are used		
Retirement contributions	4001 4003 4006 The ITR12 data are used; if no ITR12 data then IRP5 data are used	xpp	Y (it_retirement_contributions) tpnde_s
Other deductions			
Donations	4011 The assessed data are used; if no assessed data, amtdonationsallowedamtcurr is used	xcd	Y (it_deductions) tde_s
Travel claim against travel allowance	4014	xtc	Y (inc_employee) yemaltr_s, part of yem_s
Employer-provided vehicle	4048 (other than operating lease) 4050 (operating lease)	xev	Y (it_deductions) tde_s
Expenses against local and/or foreign taxable subsistence allowance	4017 (local) 4019 (foreign)	xsa xsaab	Y (inc_employee) yemalsu_s and yemabalsu_s, part of yem_s
Depreciation	4027	xde	Y (it_deductions) tde_s
Home office expenses	4028	xho	Y (it_deductions) tde_s
Amounts refunded	4042	xre	Y (it_deductions) tde_s
Allowable accountancy/ administration expense	4043	xaa	Y (it_deductions) tde_s
Legal costs	4044	xlc	Y (it_deductions) tde_s
Bad and doubtful debts	4045	xbd	Y (it_deductions) tde_s
Section 8C losses	4031	xlo	Y (it_deductions) tde_s
Expenses of holders of public office	4047	xpo	Y (inc_employee) yemalpo_s, part of yem_s
Remuneration taxed on IRP5 but complying with exemption in terms of Section 10(1)(o)(i)	4033 4041 4032	xex	Y (it_deductions) tde_s
Commission income expenditure	4016	xce	Y (inc_employee) yco_s, part of yem_s
Investments in Venture Capital Companies	4051	xvc	Y (it_deductions) tde_s

Element of PIT	Source code(s)/variable(s)	PITMOD variable name	Transformed in PITMOD? (Policy name and output variable)
Tax credits			
Rebates[24]	None		Y (it_tax_rebates) tinta_s
Medical tax credits[25]	The ITR12 data are used; if no ITR12 data then IRP5 data are used		Y (it_medical_tax_credits) tintchl_s
Medical scheme fees contributions	4005	xishl	
Medical expenses	4024	xhl	
Number of medical scheme members and dependants	meddep_1 amtmedrebatetaxcreditscalc used to calculate expected ddp	ddp ddp01	
Number of months in the scheme	amtmedrebatetaxcreditscalc, ddp and meddep_2 used to calculate a fraction	ddplv	
Actual medical scheme fees tax credit	amtmedrebatetaxcreditscalc	tintchl01	
Actual additional medical expenses tax credit	amtmedrebateexpensetaxcreditscal	tintchl02	
Disability	handicappedind	ddilv	

[24] The variable *dag* is used.

[25] The variable *dag* is also used.