Wright, Gemma Clare; Barnes, Helen; Noble, Michael; McLennan, David; Masekesa, Faith

Working Paper

Assessing the quality of the income data used in SAMOD, a South African tax-benefit microsimulation model

WIDER Working Paper, No. 2018/173

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

Suggested Citation: Wright, Gemma Clare; Barnes, Helen; Noble, Michael; McLennan, David; Masekesa, Faith (2018) : Assessing the quality of the income data used in SAMOD, a South African tax-benefit microsimulation model, WIDER Working Paper, No. 2018/173, ISBN 978-92-9256-615-9, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki,
https://doi.org/10.35188/UNU-WIDER/2018/615-9

This Version is available at:
https://hdl.handle.net/10419/211210

UNITED NATIONS UNIVERSITY
UNU-WIDER

WIDER Working Paper 2018/173

Assessing the quality of the income data used in SAMOD, a South African tax-benefit microsimulation model

Gemma Wright,[1] Helen Barnes,[1] Michael Noble,[1] David McLennan,[1] and Faith Masekesa[2]

December 2018

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

Abstract: In this paper we explore the income data in two surveys that underpin a South African tax-benefit microsimulation model. The simulated taxes and benefits using each dataset are compared with each other and with administrative data for a common time point. We explore discrepancies between the simulated and administrative data on personal income tax, with reference to the distribution of tax payers and the amount of tax simulated. Both surveys suffer from unit missing or item implausible cases for high income individuals. The paper concludes by highlighting the potential for administrative data to further enhance the quality of the model's underpinning dataset(s).

Keywords: tax-benefit microsimulation, income distribution, South Africa

JEL classification: C63, C81, H24, D31

Acknowledgements: UNU-WIDER is thanked for supporting the recent developments of SAMOD as part of the SOUTHMOD programme of work. The University of Essex is thanked for granting a license to SASPRI to use EUROMOD, and for their support of the SAMOD teams over the past decade. The European Commission is acknowledged for supporting the development of EUROMOD. An earlier version of this paper was presented at a SOUTHMOD Workshop on 13 June 2018 in Helsinki, Finland.

[1] Southern African Social Policy Research Insights (registered in the UK), corresponding author email: gemma.wright@saspri.org;
[2] Southern African Social Policy Research Institute (SASPRI) (registered in South Africa).

This study has been prepared within the UNU-WIDER project on 'SOUTHMOD – Simulating Tax and Benefit Policies for Development' which is part of the Institute's larger research project on 'The economics and politics of taxation and social protection'.

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-615-9

Typescript prepared by Ans Vehmaanperä.

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

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

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Acronyms

CDG	Care Dependency Grant
COICOP	Classification of Individual Consumption According to Purpose
CPI	Consumer Price Index
CSG	Child Support Grant
DG	Disability Grant
FCG	Foster Child Grant
GIA	Grant in Aid
ISER	Institute for Social and Economic Research, University of Essex
LCS	Living Conditions Survey
NIDS	National Income Dynamics Study
NT	National Treasury
OAG	Old Age Grant
PIT	Personal Income Tax
SAMOD	A South African tax-benefit microsimulation model
SARS	South African Revenue Service
SASPRI	Two not-for-profit organisations – Southern African Social Policy Research Insights (registered in the UK) and Southern African Social Policy Research Institute (registered in South Africa)
UIF	Unemployment Insurance Fund
UNU-WIDER	United Nations University World Institute for Development Economics Research
VAT	Value-added Tax

1 Introduction

This working paper provides an account of recent analysis to explore the impact of income data quality on simulations of taxes and benefits in South Africa.

The analysis is undertaken using a South African tax-benefit microsimulation model called 'SAMOD' which has been developed for use by government over the past decade (e.g. Wilkinson, 2009; Wright et al., 2016). SAMOD represents the first attempt to use the EUROMOD microsimulation software (e.g. Sutherland, 2001; Sutherland and Figari, 2013; Tammik, 2018)[1] in a developing country, and the EUROMOD software is now widely used in other developing country contexts.[2]

The quality of income data in any social survey that underpins a microsimulation model will inevitably impact on the simulated taxes and benefits in various ways (e.g. Ceriani et al., 2013). Income data quality issues are also of much broader interest, however, as such data inform estimates of poverty and inequality, which then feed into policy debates around poverty and inequality reduction. A number of recent studies have explored in detail the quality of income data using South African survey or Census data. These include detailed fiscal analysis using the Income and Expenditure Survey (IES) 2010/11 (Inchauste et al., 2015); spatial analysis of small area level income data using the 2011 Census of Population (Noble et al., 2013); analysis of data on wealth using the National Income Dynamics Study (NIDS) (Orthofer, 2016); in-depth analysis of the income distribution also using NIDS (Hundenborn et al., 2017) and the Post-Apartheid Labour Market Series (PALMS) dataset (Bassier and Woolard, 2018; Wittenberg, 2017). The latter four studies drew from administrative data on income tax to validate the income data in NIDS (Hundenborn et al., 2017; Orthofer, 2016) and PALMS (Bassier and Woolard, 2018; Wittenberg, 2017) respectively. There is a growing recognition of the importance of administrative data as a data resource on income, especially with reference to individuals working in the formal sector, and recently the National Treasury (NT) and the South African Revenue Service (SARS) have made anonymized personal income tax data available for research purposes (Arndt, 2018).

This paper complements earlier studies by focusing in particular on the Living Conditions Survey (LCS) 2014/15, and has an applied focus on the impact of survey income data on simulated results for the tax-benefit microsimulation model SAMOD. In terms of the structure of the paper, Section 2 provides a brief account of the tax and benefit policies that are simulated within SAMOD Version 6.6, and the two datasets that underpin the current version of SAMOD: the LCS 2014/15 (Stats SA, 2017a) and NIDS Wave 4 Version 1.1 (SALDRU, 2014). Section 3 compares the distribution of income data in the LCS and NIDS datasets, using two input ('pre-SAMOD') income concepts and simulated disposable income data. Section 4 presents SAMOD tax and benefit simulations for 2015 using the LCS and NIDS and compares these with published administrative data on benefits and taxes. Section 5 concludes with a discussion about ways in which the quality of the LCS income data could potentially be further enhanced using administrative data, informed by the issues identified here.

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

[2] See https://www.wider.unu.edu/project/southmod-simulating-tax-and-benefit-policies-development.

2 SAMOD's policies and underpinning datasets

SAMOD Version 6.6 contains systems for 2014 to 2018[3] inclusive, with policies for the taxes and benefits in June of each year. The analysis presented in this paper uses the 2015 system, to align as closely as possible with the LCS 2014/15. Table 1 summarizes the various policies that are included in the model for the 2015 system.

Table 1: Taxes and benefits that are simulated in SAMOD Version 6.6 2015 system (and June 2015 benefit amounts)

Social Assistance	
Old Age Grant	@ R1,410 per month, or R1,430 for people aged 75 or over
Disability Grant	@ R1,410 per month
Grant in Aid*	@ R330 per month
Child Support Grant	@ R330 per month
Care Dependency Grant	@ R1,410 per month
Social Allowance	
Foster Child Grant	@ R860 per month
Social Insurance	
Unemployment Insurance Fund contributions	
Taxes	
Personal Income Tax	
Value-added Tax*	

Note: * = Only simulated using LCS.

Source: Authors' compilation.

The implementation of these policies in SAMOD is described in detail in Wright et al. (2016) and so is not repeated here, except in so far as the policies could be affected by the quality of the income data in the underpinning dataset(s) (on which see especially Section 4 and Annex 1).

SAMOD Version 6.6 is underpinned by two separate datasets: the LCS 2014/15 and NIDS Wave 4 Version 1.1, both summarized in Table 2 below. The primary dataset used in this paper is the LCS 2014/15 but the NIDS Wave 4 Version 1.1 dataset provides an almost contemporaneous comparator dataset. NIDS (Wave 4 Version 1.0) performed well as an underpinning dataset for SAMOD (Wright et al., 2016)[4], and has been rigorously scrutinized with reference to external tax data (Hundenborn et al., 2017) and so is an optimal comparator dataset for the LCS 2014/15.

[3] The social benefit amounts were uprated twice in 2018 (in April and October 2018) so there are, in fact, two 2018 systems in SAMOD Version 6.6.

[4] The changes between Version 1.0 and 1.1 had limited impact on the input dataset for SAMOD.

Table 2: SAMOD Version 6.6 database description

SAMOD database	LCS 2014/15	NIDS Wave 4
Original name	Living Conditions Survey 2014/15	National Income Dynamics Study Wave 4 Version 1.1
Provider	Stats SA	SALDRU
Year of collection	2014/15	2014
Period of collection	13 October 2014 – 25 October 2015 (p. 61)	September 2014 – August 2015 (p. 26)
Income reference period	Benchmarked to April 2015 (p. 64)	Base month is November 2014*
Number of households in data	23,380	9,626**

Notes: * See the STATA do file released with NIDS Wave 4 Version 1.0, called 'Program 1d - Deflators_W4.do'.

** Having dropped non-resident and deceased households and households from Wave 3 not present in Wave 4.

Sources: For LCS 2014/15 column Stats SA (2017a). For NIDS Wave 4 column Chinhema et al. (2016).

2.1 Living Conditions Survey 2014/15

The LCS 2014/15 was undertaken by and obtained from Statistics South Africa (Stats SA). The survey was conducted between 13 October 2014 and 25 October 2015 using three data collection instruments, namely the household questionnaire, the weekly diary and the summary questionnaire.

Stats SA reports that it had adjusted the income data in two main ways (Stats SA, 2017a: 64): all income data were converted to an annual amount and benchmarked to April 2015 by inflating or deflating reported values using Consumer Price Index (CPI) data.

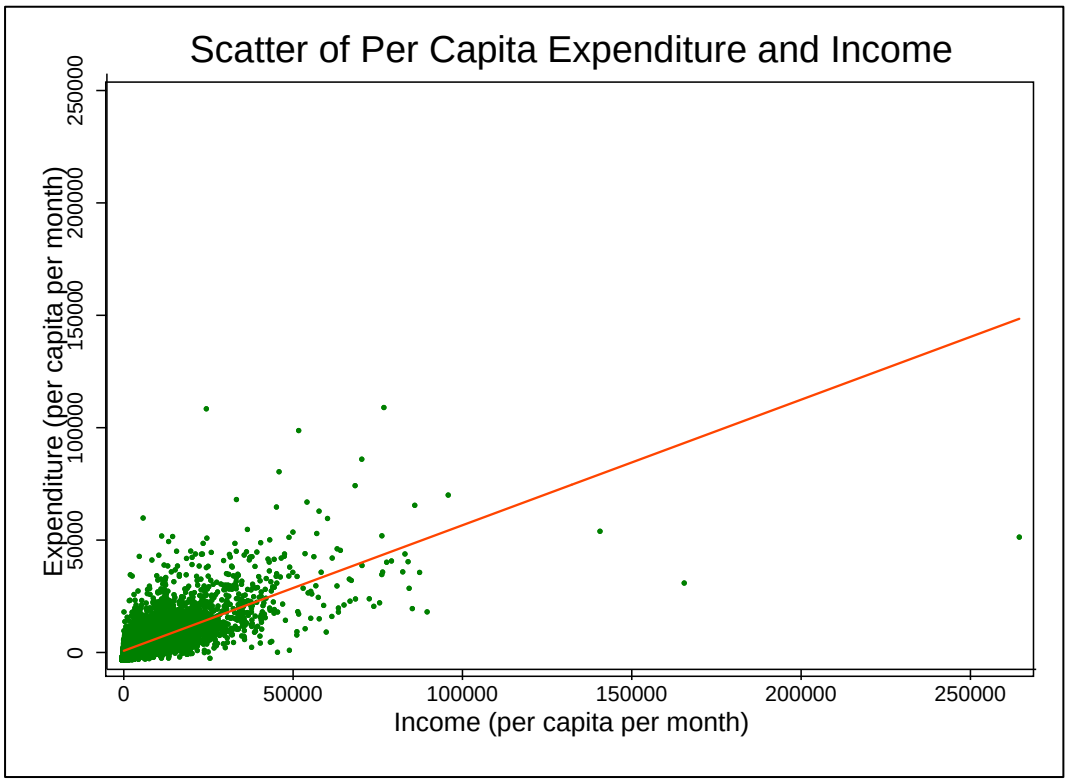
For the purposes of preparing the data for SAMOD, all monetary variables were inflated from April 2015 to June 2015 using the CPI. The final version of the LCS dataset as an input data file for SAMOD contains 23,380 households comprising 88,906 individuals.

A number of further imputations and assumptions were made in the construction of the input dataset for SAMOD Version 6.6 and these are outlined in the forthcoming Technical Note. The remainder of this section 2.1 sets out the initial steps that were undertaken to assess the quality of the income data in the LCS 2014/15.

The income information in the LCS was assessed in various ways. First, a household income was derived by summing up the values reported for each COICOP code relating to income. There were 68 households (0.3 per cent of all households) without any income information at all. Of these households, 26 (38 per cent) were in Gauteng (the next highest is 9 households or 13 per cent in North West), and 91 per cent were black African. While these households could genuinely have no or very low income, their per capita monthly expenditure (using Stats SA's composite expenditure variable) ranged from R38 to R108,974 per capita per month (see below). This could indicate implausible values for one or other of the variables. However, if both are plausible, this may suggest that households are using savings or existing credit arrangements to fund expenditure. It should also be noted that most of the very low expenditures were in rural provinces where there are former homeland areas which suggests that a combination of subsistence farming and barter may be in use. When imputed rent was deducted from household income, there were 276 households (1.2 per cent) without any income information. These households were in all provinces but mainly Gauteng (23 per cent), Limpopo (16 per cent), Northern Cape (13 per cent) and Mpumalanga (11 per cent), and 86 per cent were black African. Imputed rent is not taxable in South Africa and so has not been included in taxable income.

In order to explore the extent of implausible income values, a scatter was produced of per capita monthly expenditure and income (Figure 1).

Figure 1: Scatter of income and expenditure (Rand per capita per month values), LCS 2014/15



Source: Own calculations using LCS 2014/15.

At the top end of the income distribution there are some very high values: 236 households (1 per cent) have an income over R30,960 per capita per month (the 99th percentile). These households are located in all nine provinces but are, as expected, predominantly in the wealthier provinces of the Western Cape or Gauteng (together over 60 per cent), then KwaZulu-Natal (8 per cent) and the Eastern Cape (7 per cent). With regard to settlement type, 96 per cent of these cases are in urban formal areas.[5] The households are not large in size with 75 per cent containing 4 people or fewer. Nearly 65 per cent of households in this top percentile are white with black Africans accounting for only 25 per cent.

Employment income[6] is one of the main sources of household income, and so was explored in more detail at an individual level. There is a group of individuals who are in employment[7] but who report zero employment income for the year (less than 1 per cent of all cases, 1.3 per cent of the working age cases). There is nothing remarkable about their age, sex or population group, or province and settlement type. Of this group, half worked for a wage, 21 per cent ran a business, and 29 per cent had a job to return to. They mainly had income from letting of fixed property,

[5] Note that they are coded 1 which is urban formal but labelled urban informal – we have assumed the former.

[6] This includes COICOP codes 50110000 Household salaries and wages, 50120000 Household self-employment and business, and 50121000 Income from subsistence farming.

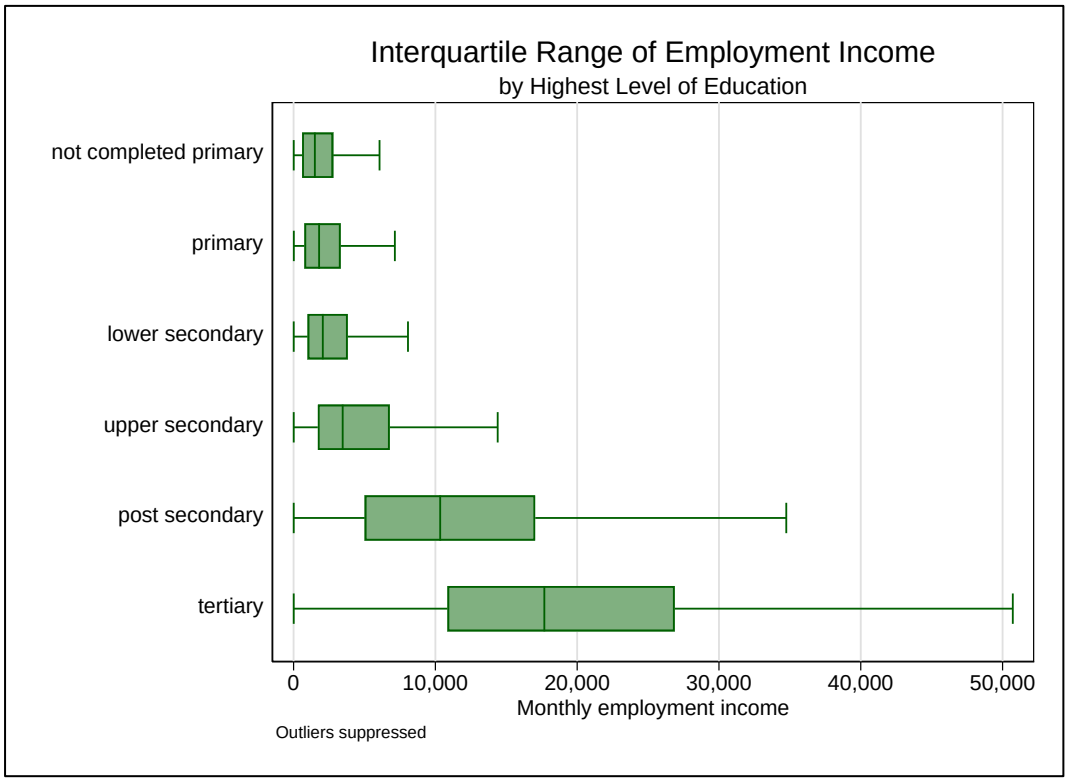
[7] Employment is defined as working for a wage/salary/commission or any payment in kind for at least one hour in the last week, or running a business for at least one hour in the last week, or having a paid job or business to return to.

social grants, alimony, palimony and other allowances and other income from individuals. The questions used to create the employment variable could capture very transient employment where pay is nominal or in kind.

On the other hand, there is a group of individuals who are not in employment but who report an employment income for the year (0.2 per cent of all cases). Some of these cases have very high employment incomes. Again, there is nothing remarkable about their age, sex or population group, or province and settlement type. Of this group, almost three quarters had salary income and the remainder had farming income. These could be people who were employed at some point over the year but not in the 7 days before the survey date, which would include seasonal workers.

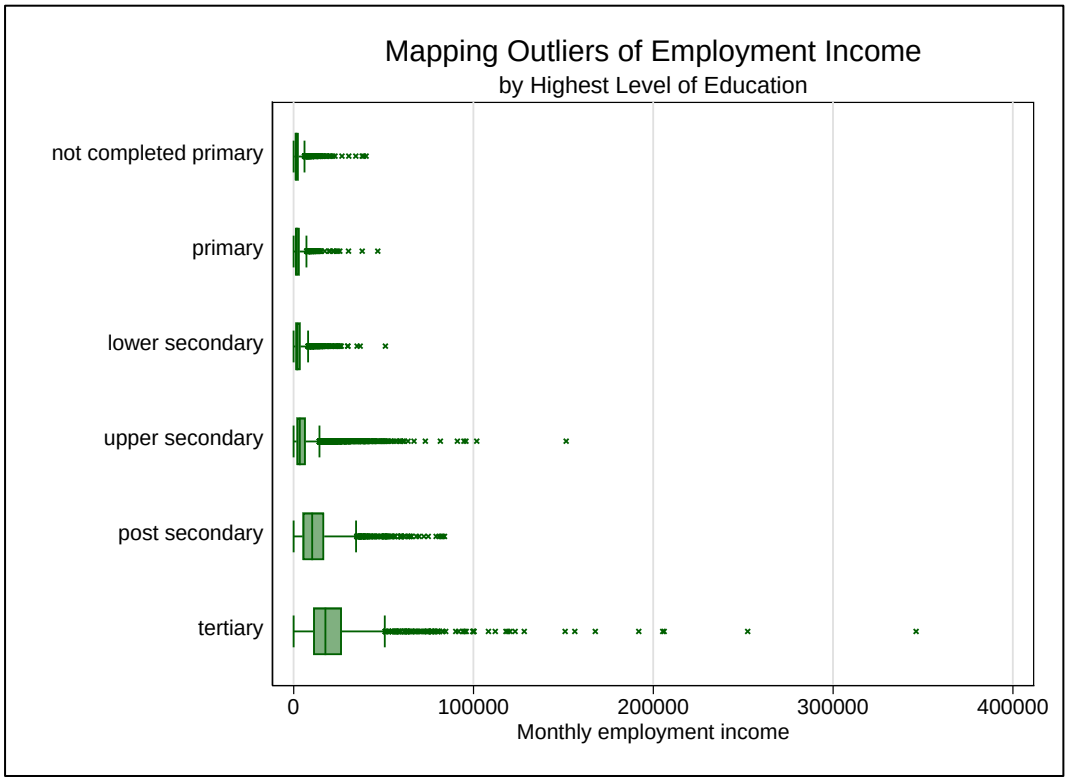
The data do not allow an occupational status variable to be constructed. However, if employment income by highest level of education is explored, the following pictures emerge (Figures 2 and 3):

Figure 2: Interquartile range of employment income by highest level of education, LCS 2014/15



Source: Own calculations using LCS 2014/15.

Figure 3: Outliers on employment income by highest level of education, LCS 2014/15



Source: Own calculations using LCS 2014/15.

Whilst there are both very high and very low incomes in the dataset, on the basis of analysis of the LCS alone, there is not any compelling evidence to suggest that the incomes are implausible. However, this is explored further in section 3 by comparing the distributions of different income concepts between the LCS and NIDS, and in section 4 with administrative data.

2.2 National Income Dynamics Study Wave 4 Version 1.1

NIDS is a national panel study carried out by the University of Cape Town (Chinhema et al., 2016). 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. Data from the fourth wave, specifically Wave 4 Version 1.1 (SALDRU, 2016), has been used as an underpinning dataset for SAMOD Version 6.6. All monetary variables were deflated from November 2014 to June 2014 using the CPI for use in SAMOD. The final dataset contains 9,626 households comprising 37,396 individuals. Adjustments made to the data are described in detail in Wright et al. (2016). These relate to NIDS Wave 4 Version 1.0 but are still applicable to Version 1.1. In addition to the data adjustments described in Wright et al. (2016), the private pension amounts of two individuals were re-categorized as lump sums as they were implausibly high for monthly amounts.

3 A comparison of SAMOD's income data using datasets derived from the LCS and NIDS

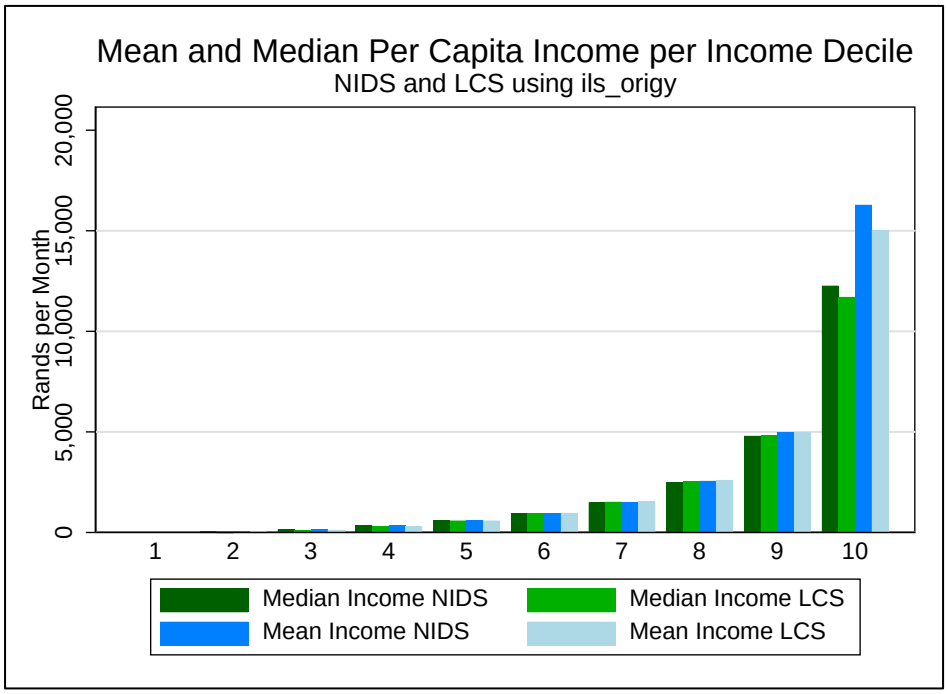
In this section, three different composite income variables are compared, with a common time point of June 2015: 'original income', income from paid employment, and simulated 'disposable income'. These were taken from the SAMOD Version 6.6 output datasets using both the LCS

2014/15 and NIDS Wave 4 Version 1.1 input datasets.[8, 9] Original income (or 'market income' or '*ils_origy*' in EUROMOD terminology) comprises gross income from employment, self-employment income, agricultural income, income from private pensions, investment income, income from private transfers, and other monetary sources of income (excluding state transfers). Income from paid employment ('*yem*' in EUROMOD terminology) does not include income from self-employment. Disposable income ('*ils_dispy*' in EUROMOD terminology) is a measure of per capita disposable income after simulation of taxes and benefits.

In order to compare the income data in the LCS and NIDS, weighted per capita income deciles of each income concept were created.

The mean and median monthly incomes for each decile of original income are shown in Figure 4 below. As can be seen, the mean and median income for deciles 1–9 are almost identical across the two datasets. However, in decile 10, NIDS has higher mean and median incomes than the LCS.

Figure 4: Mean and median per capita original income in 2015 by decile



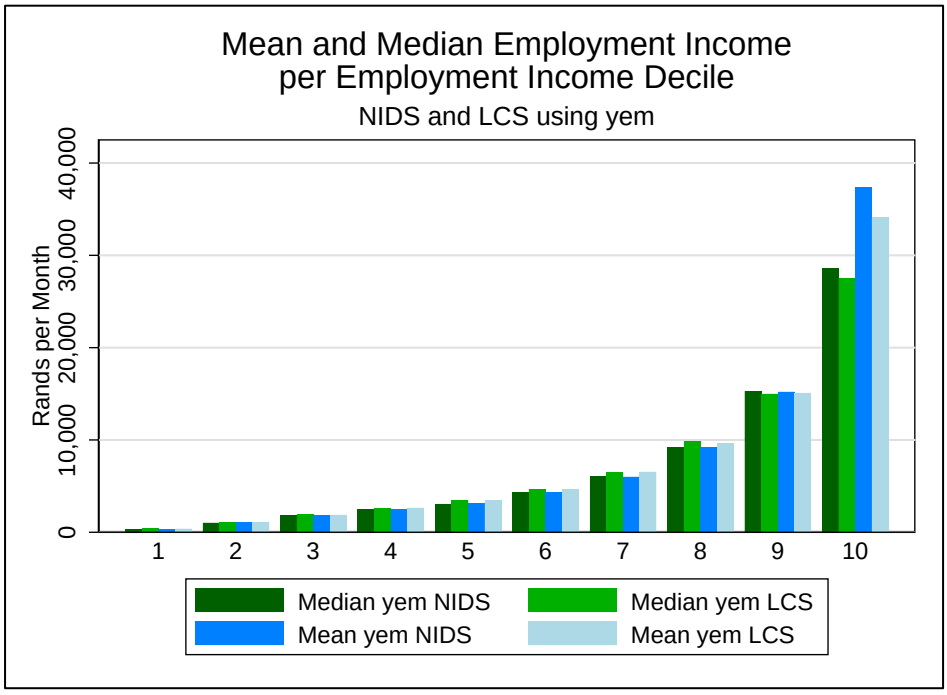
Source: Own calculations using NIDS Wave 4 Version 1.1 and LCS 2014/15.

The same figure was produced but just for income from paid employment. Figure 5 shows deciles of income from paid employment, and the mean and median values for each decile. A very similar pattern emerges to Figure 4 above.

[8] Monetary values in the NIDS dataset were uprated from June 2014 to June 2015 in SAMOD Version 6.6 using the CPI, except employment income and self-employment income which were uprated using change in average earnings.

[9] SAMOD Version 6.6 is underpinned by EUROMOD executable version 2.1.5.

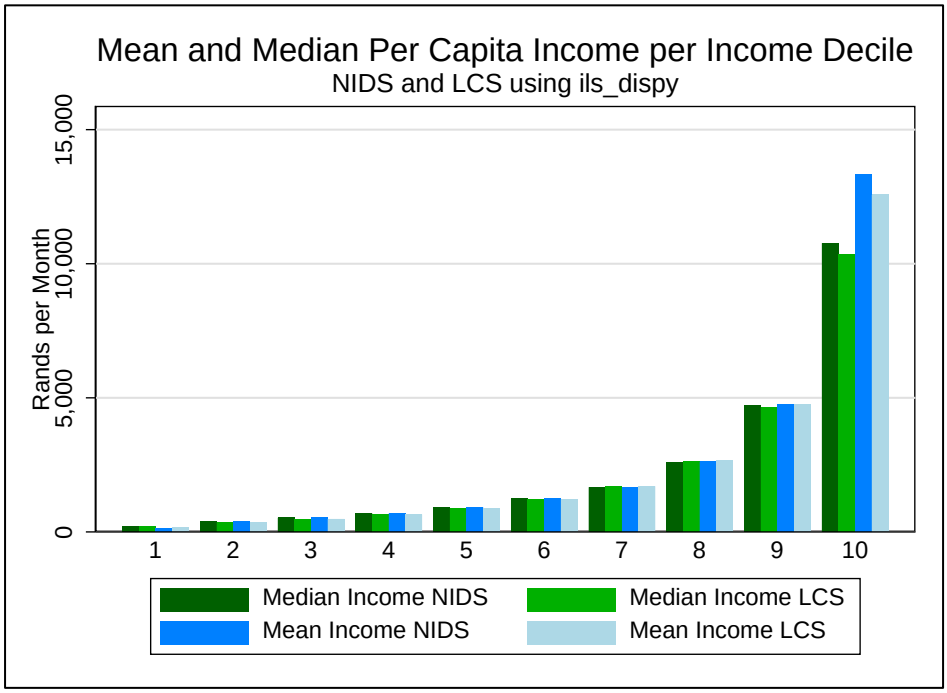
Figure 5: Mean and median employment income in 2015 by decile



Source: Own calculations using NIDS Wave 4 Version 1.1 and LCS 2014/15.

Figure 6 shows how this pattern continues through to the simulated income data from the two datasets. The figure shows the mean and median per capita disposable income by disposable income decile. Again, the NIDS data yields much higher mean and median disposable incomes than the LCS in the top decile.

Figure 6: Mean and median per capita disposable income in 2015 by decile



Source: Own calculations using NIDS Wave 4 Version 1.1 and LCS 2014/15.

Taken in isolation, the analysis presented in this section could lead one to conclude either that NIDS over-captures income in the top decile compared to the LCS, or that the LCS under-captures income in the top decile compared to NIDS. However, it could be the case that both datasets under-capture income, albeit to differing extents. For all household surveys there are concerns relating to non-response, both when a sample member fails to respond at all ('unit non-response') and when a sample member fails to respond/provide a usable response to particular items or questions ('item non-response') (Lavrakas, 2008). Such non-response can reduce statistical power and introduce bias into the survey estimates.

Stats SA (2017a) caution that the LCS 2014/15 had a lower response rate than previous household expenditure surveys, with an overall response rate of 84.9 percent. The design of the LCS weights includes a procedure for making household non-response adjustments to the base weights, but this process is 'based on the assumption that the respondent units represent both the respondent and non-respondent units' (Stats SA, 2017b: 14). However, there was a particular problem of (unit) non-response in the relatively affluent province of Gauteng (a response rate of 65.3 per cent) (Stats SA, 2017b: 5), and it is likely that this will not be adequately corrected for by the household non-response adjustments. Indeed, Stats SA highlight the impact that non-response by affluent households will have on expenditure data:

> The challenge of non-response was especially problematic in Gauteng. Given that Gauteng accounts for well over a third of all household expenditure in the country, lower responses, especially amongst high-income households in this province, means that we could see larger underreporting in various expenditure areas, most notably food, beverages and tobacco. (Stats SA, 2017a: iii).

In the same vein, it is likely that the unit missing problem will manifest itself in the income data too, and that high-income households will not have been re-introduced to a sufficient degree by the non-response adjustments that were made.

With reference to NIDS, it is known that there is concern about attrition, particularly by wealthy individuals (Finn and Leibbrandt, 2016; Rasmussen, 2017), and so both surveys will be prone to unit missing challenges.

Recent analysis by Hundenborn et al. (2017) has identified that NIDS income data compares well with a large anonymized sample of tax records, except for the highest income decile. On the basis of Hundenborn's finding and the analysis presented here which shows that the incomes of deciles 1–9 are very similar for both NIDS and the LCS (Figures 4-6), it can be hypothesized that—or those with taxable incomes—the unit missing problem mainly lies in the tenth decile for the LCS too, and possibly to an even greater extent than for NIDS.

In order to unpack these issues further, in the next section the simulated taxes and benefits using data derived using the LCS and NIDS are compared with external administrative data on tax revenue and benefits paid.

4 External validation of simulated benefits and taxes

This section presents simulated benefits and taxes that were produced using SAMOD Version 6.6 for a June 2015 timepoint using the input datasets derived from NIDS Wave 4 Version 1.1 and the LCS 2014/15 and compares them with published administrative data sources.

All but two of SAMOD's simulated policies could be affected by the quality of the income data in the underpinning dataset. The two exceptions are Value-added Tax (VAT) as this is implemented using household expenditure data rather than income data; and the Foster Child Grant as this is a social allowance and therefore has no means-test and so will be unaffected by the quality of the income data. For all other policies, the quality of the income data in SAMOD's underpinning dataset will affect the simulated results in various ways including the number of individuals identified within SAMOD as being eligible for certain benefits and eligible to pay Personal Income Tax (PIT); the simulated amount of government expenditure required to pay benefits to eligible individuals; and the size of the anticipated revenue to government in the form of PIT. Annex 1 provides further details about the survey income variables that are used in SAMOD's tax and benefit policies and how they could impact on the simulated results.

Regarding numbers of grant beneficiaries, Table 3 compares the number of reported beneficiaries in June 2015 (SASSA, 2015) with the simulated numbers of eligible beneficiaries for each grant using NIDS and the LCS. Taking the OAG as an example, SASSA report that there were 3.1 million recipients of the OAG in June 2015; using SAMOD it is estimated that there were 3.8 million eligible older people using NIDS, or 3.7 million eligible older people using the LCS.

Table 3: Reported and simulated grant beneficiary numbers in 2015

Grant	Reported (SASSA, 2015)	NIDS wave 4 version 1.1			LCS 2014/15		
		SAMOD simulated	% captured (simulated/ reported)	Take-up (reported/ simulated)	SAMOD simulated	% captured (simulated/ reported)	Take-up (reported/ simulated)
CDG	127,869	148,930	116	86	110,113	86	116
FCG	519,031	549,314	106	94	612,511	118	85
CSG	11,792,596	14,854,041	126	79	14,489,39517	123	81
OAG	3,114,729	3,841,262	123	81	3,726,687	120	84
DG	1,106,425	1,259,570	114	88	1,207,564	109	92
GIA	119,541	/	/	/	122,036	102	98

Notes: The NIDS Wave 4 results shown in this table for the number of beneficiaries are slightly different from those reported in Wright et al. (2016) because the latter used NIDS Wave 4 version 1.0 and an earlier version of SAMOD, but the take-up rates are unaffected.

Source: Own calculations using SAMOD Version 6.6 and reported figures from SASSA (2015).

The reported figures from administrative data relate to actual payment of social grants and will be subject to errors of inclusion and errors of exclusion, whereas the simulated figures from the NIDS or LCS datasets are estimates of eligibility for the grant. However, this does enable estimates of take-up to be calculated. Using the OAG as an example again, the estimated take-up rates for OAG are 81 per cent using NIDS, and 84 per cent using the LCS. Overall, the simulated numbers of beneficiaries—and the estimated take-up rates—using NIDS and the LCS are similar for the main benefits (CSG, OAG and DG).

Table 4 presents expenditure on social grants, comparing reported expenditure with simulated expenditure. Again, NIDS and the LCS produce similar results for the main benefits (CSG, OAG and DG).

Table 4: Reported and simulated cost of each grant in 2015, assuming full take-up

Grant	Repor-ted (Rm) (NT, 2016)	NIDS wave 4 version 1.1			LCS 2014/15		
		SAMOD simulated (Rm)	% captured (simulated / reported)	Expenditure take-up (reported/ simulated)	SAMOD simulated (Rm)	% captured (simulated / reported)	Expendi-ture take-up (reported/ simulated)
CDG	2,431	2,520	104	96	1,863	77	130
FCG	5,480	5,669	103	97	6,321	115	87
CSG	47,459	58,822	124	81	57,378	121	83
OAG	53,274	62,860	118	85	60,565	114	88
DG	19,298	21,135	110	91	20,020	104	96

Note: Official sources only report beneficiary numbers for GIA and so results for GIA costs are not included in this table.

Source: Own calculations using SAMOD Version 6.6 and revised estimates for 2015/16 from NT (2016: 64).

Regarding direct taxes, Table 5 provides information about the number of personal income tax payers in 2015. The NT estimate of registered tax payers (just over 7 million) is higher than two figures published by NT and SARS: 6.55 million individuals expected to submit returns (Row 2) and 5.37 million assessed individuals (Row 3). Regarding the simulated numbers of tax payers, NIDS yields 5.41 million tax payers, compared to a slightly higher 5.49 million using the LCS 2014/15. The simulated number of tax payers using NIDS and the LCS are therefore similar to each other and are both closest to NT and SARS 'assessed individuals' figure of 5.37 million tax payers, with NIDS simulating 101 per cent, and LCS simulating 102 per cent of this NT and SARS figure.

The interpretation of Table 5 is not straightforward as the simulated figures do not precisely correspond to concepts used in the published administrative data. Importantly, the simulated number of tax payers in SAMOD reflects the tax liability for the year, given the incomes of individuals for that year. In accountancy terms, this is referred to as an 'accrual' basis.[10] In contrast, the administrative data is published on a 'cash-flow' (or in this instance 'taxpayer-flow') basis and so none of the reported figures precisely correspond to what is simulated by SAMOD. The NT and SARS 'expected to submit' figure (6.55 million) comprises all tax payers who have been assessed for the tax year, plus taxpayers with an 'active' status who were assessed in any of the two previous years (NT and SARS, 2017). The NT and SARS 'assessed individuals' figure (5.37 million) comprises only those tax payers whose tax submission has been assessed, a figure which changes in later years as more outstanding returns are submitted and processed (NT and SARS, 2017). Furthermore, both of the NT and SARS figures exclude certain tax payers who were employed by a single employer for the year of assessment and paid tax through PAYE and have a gross employment income/salary below R350,000 (the submission taxable income threshold).[11] Thus there is a sizeable group of tax payers that is missing from the NT and SARS figures. Assessed individuals may include those paying arrears of taxes and fines or those receiving refunds of tax from previous years, and will not include those with tax liabilities for the current year which are not due until subsequent years.

[10] The concepts of accrual and cash flow reporting are explained in more detail in the forthcoming Technical Note.

[11] 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).

The NT estimate of registered tax payers (7.02 million) is closer to the SAMOD simulations in definitional terms, though the figure may well include individuals on the register[12] but not active for a particular tax year.

Table 5: Reported and simulated number of tax payers in 2015

	Reported	NIDS wave 4 version 1.1		LCS 2014/15	
		SAMOD simulated	% captured (simulated/reported)	SAMOD simulated	% captured (simulated/reported)
1. NT	7,024,199	5,413,740	77	5,487,671	78
2. NT and SARS (i)	6,554,174	5,413,740	83	5,487,671	84
3. NT and SARS (ii)	5,370,717	5,413,740	101	5,487,671	102

Source: Own calculations using SAMOD Version 6.6 and reported figures:

1. NT (2015: 48) – Estimates of registered individuals with taxable income above taxable income tax threshold, 2015/16 (reported as 2014/15 but this appears to be a typographical error based on publications from previous years).

2. NT and SARS (2017: 36) – Number of individuals expected to submit returns, 31 March 2015.

3. NT and SARS (2017: 36) – Number of assessed individuals, 31 March 2015 (2015 tax year).

Regarding tax revenue, the same definitional challenges exist when making comparisons with published figures. So, for example, the NT and SARS figure will include PIT from those PAYE cases with taxable income over R350,000 or from those whose tax submission has been assessed. It will not include tax collected through PAYE unless the individual is required to submit a tax return. For any particular year, the NT and SARS figure increases and is updated retrospectively as more outstanding returns are submitted and processed. The NT figure is the reported receipts for the tax year in question and so will include all PAYE cases. However, it will also include, for example, arrears from previous years.

In addition to differences between the two figures from administrative sources, the NIDS and LCS simulated figures diverge. Using NIDS, R229 billion of PIT is simulated, compared to R187 billion using the LCS (Table 6). The amount of PIT simulated using NIDS is 59 per cent of the NT figure and 79 per cent of the SARS figure, compared to just 48 per cent or 65 per cent respectively using the LCS.

[12] Everyone formally employed, regardless of their tax liability, must be registered for PIT (NT and SARS, 2017).

Table 6: Reported and simulated revenue from personal income tax in 2015

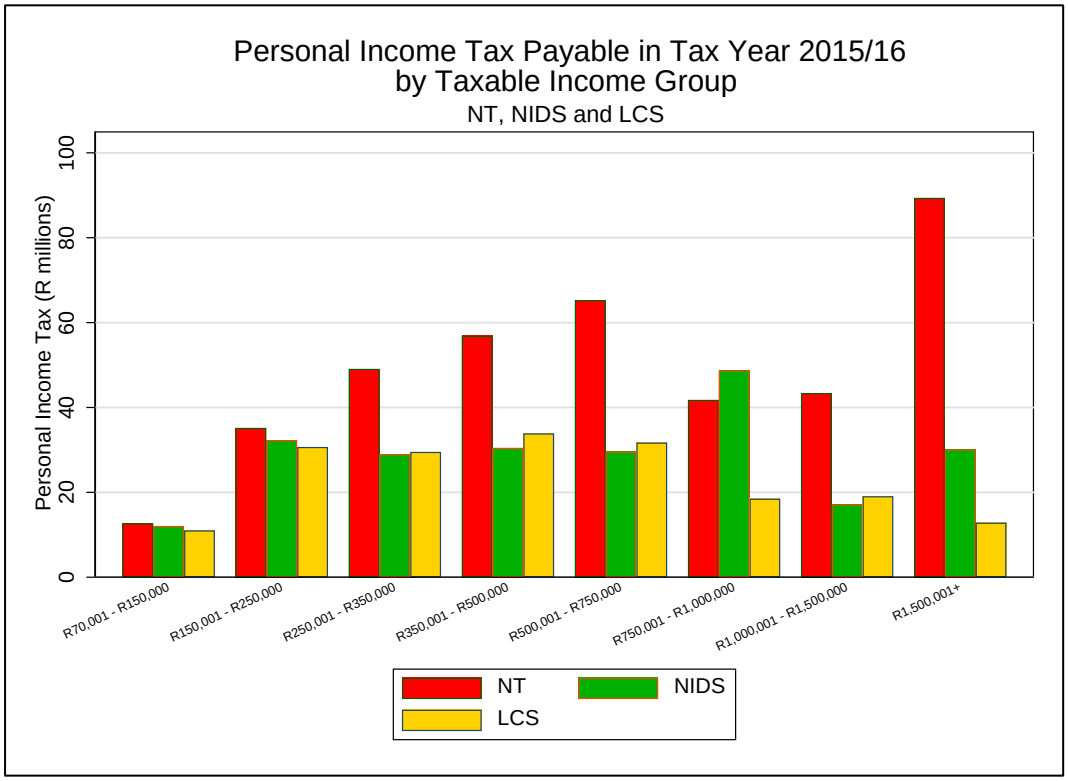
	Reported (Rm)	NIDS wave 4 version 1.1		LCS 2014/15	
		SAMOD simulated (Rm)	% captured (simulated/reported)	SAMOD simulated (Rm)	% captured (simulated/reported)
1. NT	388,102	229,375	59	187,388	48
2. NT and SARS	288,873	229,375	79	187,388	65

Source: Authors' calculations using SAMOD Version 6.6 and reported figures:

1. NT (2017: 43) – Budget revenue, personal income tax, outcome for 2015/16.
2. NT and SARS (2017: 37) – Tax assessed, 2015 tax year.

In order to explore these findings about PIT in more detail, comparisons were made with published information on 'taxable income group'[13] with respect to estimates of the number of taxpayers and the amount of income tax payable for the 2015/16 tax year (NT, 2015: 48). It was possible to produce a set of taxable income groups in both the LCS and NIDS output datasets using the simulated variable *ttb_s* (taxable income).[14]

Figure 7: A comparison of the amount of PIT paid by taxable income group using NT administrative data and NIDS and LCS data, 2015/16



Note: NT data is derived from administrative sources and is an estimate of the amount of PIT. NIDS and LCS data are derived from SAMOD simulations. The taxable income groups presented are those used by NT for reporting purposes. Source: NT: NT (2015). LCS and NIDS: Own calculations using SAMOD V6.6.

[13] '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.

[14] To which monthly lump sum income was added.

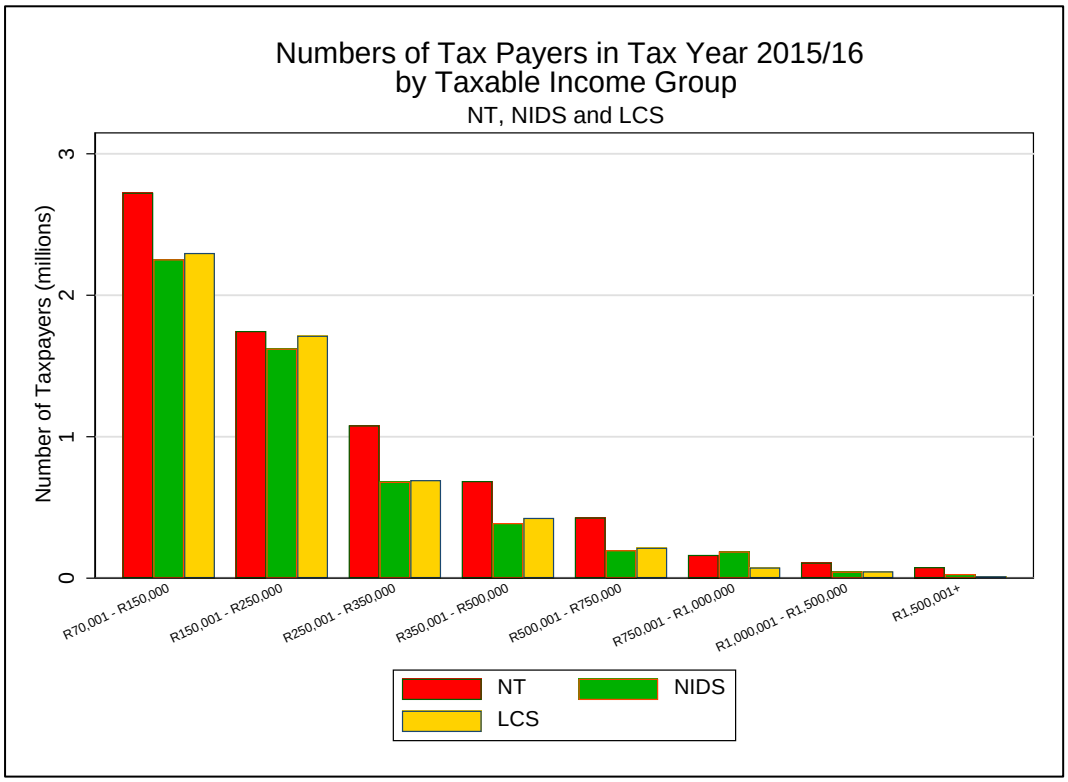
Figure 7 shows the amount of tax paid by taxable income group, comparing reported amounts from NT (in red)[15] with simulated amounts using SAMOD (the variable *tin_s*) and based on NIDS (in green) and the LCS (in yellow) and assuming full compliance. Comparing the LCS and NIDS first, the simulated amounts are very similar, particularly for the lower taxable income groups. The largest differences between the LCS and NIDS simulated amounts occur in the R750,001–R1,000,000 and R1,500,000+ groups where NIDS simulates a higher amount of tax. Comparing the LCS and NIDS simulated amounts with the NT reported amounts, it is evident that the LCS simulates too little tax for all taxable income groups, though the difference between the LCS simulations and reported amounts is negligible for the lowest two groups (R70,001–R150,000 and R150,001–R250,000). The same is broadly true for NIDS, the exception being the third highest group (R750,001–R1,000,000) where a higher amount of tax is simulated using NIDS than is reported by the NT. The largest discrepancy between the NIDS and LCS simulated amounts and the NT reported amounts is for the final group (R1,500,000+) where the NIDS and LCS amounts are much lower than the NT amounts.

The findings for the highest taxable income group can be interpreted in a number of ways: both surveys might have insufficient numbers of wealthy individuals in their enumerated samples (this would be categorized as a unit missing problem, for individuals with high taxable incomes); alternatively, wealthy individuals in the surveys could be under-reporting their taxable income in the surveys (this would be categorized as an item missing or item implausible data problem, for individuals with high taxable incomes); or both situations could be occurring in combination. Regarding the third highest group, the higher NIDS figure could simply be a function of a few individuals in this taxable income group in NIDS having weights that are too large, or could signal under-reporting of income by NIDS respondents who ought to be in the highest taxable income group.

Figure 8 shows the number of tax payers reported by NT (in red) and simulated using SAMOD and NIDS (in green) and the LCS (in yellow).

[15] It was not possible to obtain published outcome information on either number of taxpayers or tax payable by taxable income group, and therefore the total PIT amount relating to Figure 7 (R394 billion) is based on an estimate and differs slightly from the amount presented in Table 6 (R388 billion), which is the outcome reported for the 2015/16 tax year. It may therefore be the case that some of the discrepancies in Figure 7 are overstated.

Figure 8: A comparison of the number of tax payers by taxable income group using NT administrative data and NIDS and LCS data, 2015/16



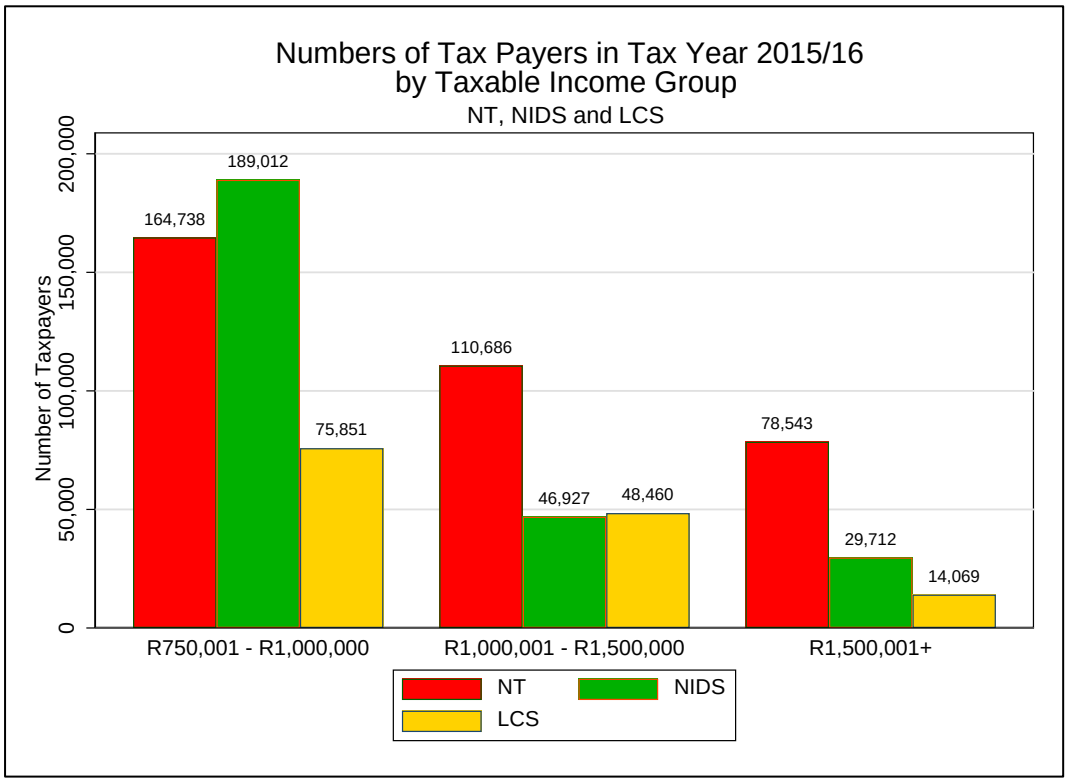
Note: NT data is derived from administrative sources and is an estimate of the number of registered tax payers. NIDS and LCS data are derived from SAMOD simulations. The taxable income groups presented are those used by NT for reporting purposes.

Source: NT: NT (2015). LCS and NIDS: Own calculations using SAMOD V6.6.

From Figure 8 it can be observed that in general SAMOD simulates fewer tax payers than NT in all taxable income groups, whether using NIDS or the LCS. The extent of the discrepancy is fairly small in the second taxable income group (R150,001–R250,000), particularly with regard to the LCS simulations which yield a very similar figure to the NT estimate. In the third highest group (R750,001–R1,000,000), NIDS has a slightly higher number of tax payers than reported by NT which in turn is higher than the LCS estimate.

Figure 9 shows the number of tax payers in the highest three taxable income groups which, due to small numbers, cannot be seen clearly on Figure 8. This highlights the mismatch between the administrative data and SAMOD simulations for all three groups, with the LCS and NIDS simulating less than half of the NT estimates in each group. The exception is the NIDS simulated figure for the third highest group discussed above. Figure 9 also reveals a discrepancy between NIDS and the LCS in the highest band, with the LCS simulating approximately half of the tax payers that are simulated using NIDS.

Figure 9: A comparison of the number of tax payers in the highest three taxable income groups using NT administrative data and NIDS and LCS data, 2015/16



Source: NT: NT (2015). LCS and NIDS: Own calculations using SAMOD V6.6.

What can be deduced from these findings? If it is assumed that there is full compliance, that is assuming that the administrative data correctly reflects the number of tax payers for each taxable income group, then it would seem that the two surveys fail to capture individuals in all taxable income groups. The absence in the surveys of a relatively small number of individuals with high taxable incomes (unit missing) or under-disclosure of their incomes in surveys (item missing/implausible) will have—and evidently does have—a large impact on the total amount of PIT simulated using these surveys, as demonstrated in Figure 7. However, the impact on the tax take of an absent or under-reporting lower income individual is not as great, as evidenced above, where the number of taxpayers in the lower taxable income groups simulated by the LCS and NIDS is lower than reported by NT (Figure 8), yet the amount of tax paid by these groups is only marginally less in the SAMOD simulations compared to the NT reported amount (Figure 7).

5 Concluding remarks

This paper presents analysis to explore the quality of the income data in the LCS 2014/15. Section 2 provided an account of various internal tests that were conducted on the income data. Section 3 compared the distribution of different income concepts for the LCS and an almost contemporaneous NIDS dataset. The analysis presented in Section 3 demonstrates that a divergence occurs at the top income decile between the NIDS and LCS datasets. This divergence occurs in the cleaned and prepared market income variable and employment income variable for SAMOD (Figures 4 and 5) and is carried through to the simulated income data on disposable income (Figure 6).

In Section 4, SAMOD's simulated taxes and benefits were compared using LCS and NIDS data, and it was demonstrated that the number of grant beneficiaries is broadly similar across the two surveys, particularly for the main grants (CSG, OAG and DG). Regarding direct taxes, the number of tax payers identified in SAMOD is also similar using LCS and NIDS data (Table 5), and corresponds very closely to the NT and SARS figure of assessed individuals. However, there are important definitional differences between SAMOD simulations and the NT and SARS figure and so this observation should be treated with caution—other comparators from administrative data correspond less well with the SAMOD simulations. The main divergence occurs in the simulation of PIT, with SAMOD simulating less PIT with the LCS dataset than with NIDS, and both surveys simulating less PIT than reported figures from NT, and NT and SARS (Table 6). In order to explore why this divergence might be occurring, comparisons were made between SAMOD's simulated figures using LCS and NIDS, and published data for number of tax payers and taxable income by taxable income group (a categorization used for reporting purposes rather than the income tax brackets used in the calculation of PIT).

The taxable income group analysis can be interpreted in a number of ways (Figures 7 and 8). In terms of simulated taxes, it is most likely that the analysis demonstrates that both surveys suffer from unit missing data for high income individuals (R1 million and over): SAMOD simulates far fewer tax payers in the highest taxable income group, whether using the LCS or NIDS, than reported by NT (though there are definitional differences). Regarding the lowest taxable income groups, although SAMOD simulates fewer tax payers in these groups, whether using the LCS or NIDS, this does not result in a large discrepancy in taxes paid between the SAMOD simulations and NT reported amounts as they are low-income individuals paying small amounts of tax.

There are many ways in which this analysis could be developed further to enhance the quality of the income data in the LCS. Given the better performance of NIDS with regard to simulating PIT revenue with SAMOD, one option might have been to adjust incomes upwards in the LCS to correspond to NIDS. However, this does not seem worthwhile, even if restricted to the top decile, as both NIDS and the LCS fall short of the administrative data on PIT and almost certainly both suffer from the challenge of unit missing data for high income individuals. Regarding NIDS, Rasmussen has commented:

> 'Given the small numbers of high income earners in the dataset and that NIDS is a longitudinal survey (in which one may expect high income households to either be under sampled or have higher rates of attrition), it is important to view the results based on the top end of the income distribution with a degree of scepticism. These earners would perhaps be better captured by tax data.' (Rasmussen, 2017: 4)

Rasmussen's point that high earners in NIDS could be better captured by tax data is well taken and also applies, perhaps more so, to the LCS.

There are two main ways in which administrative tax records could be used to improve the quality of income data in social surveys. The first way is by using aggregate tax data on numbers of taxpayers and the amount of taxes they pay. Such aggregate data (used to some extent in this paper) could help to determine whether and to what extent there is a unit missing problem. It could then be used to assist in the re-weighting of cases to address this issue. A further refinement would be to obtain administrative data by sub-group (e.g. residence of taxpayer, gender of taxpayer and type of employment) would allow a more accurate diagnosis of the extent of unit missing information within the survey as well as a more sophisticated re-weighting approach.

The second way in which tax data could be useful is by some form of data linkage. Information from administrative tax records could be linked to specific cases and thus could enable issues of

item missing or item implausible to be identified and addressed. There are various data matching techniques that could be explored including case-by-case matching in a secure setting, hierarchical matching, and probabilistic matching methods such as propensity score matching (see for example Jenkins et al., 2007; Harron, 2016). Direct matching would reveal the extent to which item implausible or item missing cases were present and would enable corrections to be made.

This paper has highlighted the definitional challenges of comparing like with like when using aggregate tax records. As noted, tax information is often presented in published data on a cash flow basis rather than an accrual basis, whereas SAMOD's estimates are accrual-based. In order to use tax records to greater effect it will be important to try to obtain accrual-based figures from administrative sources that cover the same period as the modelled estimates. The NT, and NT and SARS figures presented both had shortcomings in terms of the specific groups they captured and the means by which the figures were derived. The ideal administrative data would be based on outcomes rather than estimates, and for all taxpayers rather than, for example, only those with income above the submission threshold.

References

Arndt, C. (2018). 'New data, new approaches and new evidence: a policy synthesis'. *South African Journal of Economics,* 86(1): 167–78.

Bassier, I. and Woolard, I. (2018). 'Exclusive growth: Rapidly increasing top incomes amidsts low national growth in South Africa'. REDI3x3 Working Paper No.47. Cape Town: Research Project on Employment, Income Distribution and Inclusive Growth, SALDRU, University of Cape Town. Available at: http://www.redi3x3.org/sites/default/files/Bassier%20%26%20Woolard%202018%20REDI3x3%20Working%20Paper%2047%20Top%20incomes.pdf.

Ceriani, L., Fiorio, C.V., and Gigliarano, C. (2013). 'The importance of choosing the data set for tax-benefit analysis'. *International Journal of Microsimulation,* 6(1): 86–121.

Chinhema, M., Brophy, T., Brown, M., Leibbrandt, M., Mlatsheni, C.. and Woolard, I. (eds) (2016). *National Income Dynamics Study Panel User Manual.* Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town.

Elliot, M., O'Hara, K., Raab, C., O'Keefe, C.M., Mackey, E., Dibben, C., Gowans, H., Purdham, K. and McCullagh, K. (2018). 'Functional Anonymisation: Personal Data and the Data Environment'. *Computer Law and Security Review,* 34(2): 204–21.

Finn, A. and Leibbrandt, M. (2016). 'The dynamics of poverty in the first four waves of NIDS'. SALDRU Working Paper No.174/NIDS Discussion Paper 2016/1. Cape Town: SALDRU, University of Cape Town.

Harron, K. (2016). *Introduction to Data Linkage.* Edited by E. Mackay and M. Elliot. Colchester: Administrative Data Research Network, University of Essex.

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

Inchauste, G., Lustig, N., Maboshe, M., Purfield, C. and Woolard, I. (2015). 'The distributional impact of fiscal policy in South Africa'. CEQ Working Paper No.29. New Orleans, LA: Commitment to Equity project, Tulane University.

Jenkins, S.P., Lynn, P., Jackle, A. and Sala, E. (2007). 'The feasibility of linking household survey and administrative record data: new evidence for Britain'. *International Journal of Social Research Methodology*. 11(1): 29–43.

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

McLennan, D. (2018). *Data quality issues in administrative data*. Edited by E. Mackay and M. Elliot. Colchester: Administrative Data Research Network, University of Essex.

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

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

National Treasury (2016). 'Budget Review 2016'. Pretoria: National Treasury South Africa.

National Treasury (2017). 'Budget Review 2017'. Pretoria: National Treasury South Africa.

National Treasury and South African Revenue Service (NT and SARS) (2017). *Tax Statistics 2017*. Pretoria: National Treasury and South African Revenue Service.

Noble, M., Zembe, W., Wright, G. and Avenell, D. (2013). *Multiple Deprivation and Income Poverty at Small Area Level in South Africa in 2011*. Cape Town: SASPRI.

Orthofer, A. (2016). 'Wealth inequality in South Africa: evidence from survey and tax data'. REDI3x3 Working Paper 15. Cape Town: Research Project on Employment, Income Distribution and Inclusive Growth, 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.

South African Social Security Agency (SASSA) (2015). 'SASSA Fact Sheet: Issue no 6 of 2015 – 30 June 2015'. Pretoria: SASSA.

Southern Africa Labour and Development Research Unit (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.

Statistics South Africa (Stats SA) (2011). 'Living Conditions of Households in SA 2008/2009'. Statistical Release P0310. Pretoria: Statistics South Africa.

Statistics South Africa (Stats SA) (2015a). 'Consumer Price Index August 2015'. Statistical Release P0141. Pretoria: Statistics South Africa.

Statistics South Africa (Stats SA) (2015b). 'Quarterly Employment Statistics June 2015'. Statistical Release P0277. Pretoria: Statistics South Africa.

Statistics South Africa (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.

Statistics South Africa (Stats SA) (2017b). 'Living Conditions Survey 2014/2015: Metadata'. Statistical Release P0310 (2017). Pretoria: Statistics South Africa.

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

Sutherland, H., and F. Figari (2013). 'EUROMOD: the European Union tax-benefit microsimulation model'. *International Journal of Microsimulation*, 6(1): 4–26.

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

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 No. 212. Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town.

Wright, G., Noble, M., Barnes, H., McLennan, D. and Mpike, M. (2016). 'SAMOD, a South African tax-benefit microsimulation model: recent developments'. WIDER Working Paper No. 2016/115. Helsinki: UNU-WIDER.

Wright, G., Noble, M., Dinbabo, M., Ntshongwana, P., Wilkinson, K., and Le Roux, P. (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.

Annex 1 Use of self-reported income data in SAMOD's input datasets

SAMOD Version 6.6 is underpinned by the LCS 2014/15 (Stats SA, 2017a) and NIDS Wave 4 Version 1.1 (SALDRU, 2014). The quality of the income data in these surveys will have a direct impact on most of the simulated taxes and benefits in SAMOD. This Annex provides more detail about the self-reported income variables that are used in the simulation of South Africa's tax and benefit system using SAMOD.

Regarding the benefits, the Old Age Grant (OAG), Disability Grant (DG), Grant in Aid (GIA), and Care Dependency Grant (CDG) all have the same means-tests of R64,680 per year for single people, and R129,360 for couples, as at June 2015. The OAG is payable to those aged 60 and over while the non-overlapping DG is payable to people aged 18–59 with a work-limiting disability, with GIA being a top-up for disabled adults in need of constant care. The CDG is payable for disabled children in need of constant care. The means-test for the Child Support Grant (CSG) in June 2015 was R39,600 per year for single caregivers, and R79,200 per year for caregivers with a spouse. Thus, if there is missing income data that if present would have been taken into account within a means-test, this would mean that SAMOD would simulate too many eligible beneficiaries. Conversely, if incomes are erroneously over-reported then SAMOD would simulate too few eligible beneficiaries for means-tested benefits.

The incomes that are taken into account for means-tested benefits are in fact made up of various components, most of which are obtained from the self-reported survey data. Table A1 lists the various income and expenditure categories that are taken into account in the means-tests, and provides the variable name as used in SAMOD, as well as specifying whether the variables are obtained from the survey data ('Survey') or generated on-model ('SAMOD').

Table A1: Types of income or expenditure that are taken into account for means-tested benefits in SAMOD

Type	Obtained from self-reported survey data or simulated on-model?	Variable name in SAMOD
Income from employment	Survey	yem
Income from self-employment	Survey	yse
Income from property	Survey	ypr
Income from private pension	Survey	ypp
Income from compensation	Survey	ycm
Income from investment interest	Survey	yiyit
Expenditure on health insurance by employee	Survey	xishl
Expenditure on pension contributions	Survey	xpp
PIT payable	SAMOD	tin_s
Employee contribution to UIF	SAMOD	tscee_s

Source: Authors' compilation using SAMOD Version 6.6.

The composite income list comprising means-testable income (*ils_social_grants*) is calculated for the purposes of the means-tested benefits as follows:

Ils_social_grants=yem+yse+ypr+ypp+ycm+yiyit-xishl-xpp-tin_s-tscee_s

Regarding the Unemployment Insurance Fund (UIF), employer and employee contributions are each simulated in SAMOD at the rate of 1 per cent of gross salary up to a threshold salary of R14,872 per month. Thus, if employment income is under-reported, SAMOD will under-estimate

the amount of UIF payable by employees and employers; and if employment income is erroneously over-reported this would result in over-estimates of UIF revenue.

The impact of missing or erroneous income data on the Personal Income Tax (PIT) is more complex to elucidate, but broadly speaking over-reported income would cause SAMOD to simulate too much PIT, and missing or under-reported income data would cause SAMOD to simulate too little PIT.

SAMOD uses self-reported survey data for various income types as part of the PIT policy. In order to better appreciate the potential impact on simulations of PIT of survey income data, Table A2 below 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 survey data using self-reported information ('Survey'), or are generated on-model ('SAMOD').

Table A2: Types of income or expenditure that are taken into account or generated in the PIT policy in SAMOD

Type	Obtained from self-reported survey data or simulated on-model?	Variable name in SAMOD
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
Tax payable on income from interest payments	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
Tax payable on retirement related lump sum	SAMOD	ttaoy_s
Income from employment related lump sum	Survey	ysv
Tax payable on employment related lump sum	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 6.6.

The PIT calculation in SAMOD can be summarized as follows:

Income tax payable = Tax payable on (general taxable income + 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_taxable[16] *+ ttaiy_s - ttapn_s)*[17] *+ tinkt_s - tinta_s - tintchl_s*

The composite variable *tinkt_s* is the sum of tax on retirement related lump sums (*ttaoy_s*) and employment related lump sums (*ttasv_s*). 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 R500,000 per year. The excess lump sum amounts are taxed at 18 per cent for amounts less than or equal to R200,000 per year; 27 per cent for amounts of R200,001 to R550,000 per year; and 36 per cent for amounts of R550,001 and over per year.

The general tax rebate (*tinta_s*) is simulated on-model within SAMOD for 2015 at R13,257 per year, with an additional rebate of R7,407 for those aged 65 and over, and a further R2,466 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 excess medical scheme fees tax credit relating to additional fees paid, 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 2015 at 18 per cent for the first R181,900 per year; 26 per cent for R181,901 to R284,100; 31 per cent for R284,101 to R393,200; 36 per cent for R393,201 to R550,100; 39 per cent for R550,101 to R701,300; and 41 per cent for amounts above R701,300.

In summary, the means-tested benefit policies will be affected by the quality of six income variables and two expenditure variables in the survey data; and the PIT policy will be affected by the quality of eight income variables and four expenditure variables.

Lastly, the quality of any of these variables will impact in different ways on different parts of the income distribution. For example, using the same LCS 2014/15 dataset, Statistics South Africa report that income from work (*yem* and *yse* in SAMOD terminology*)* accounts for just 14 per cent of annual household income in the lowest income per capita decile, and more than 70 per cent of annual household income in deciles 6-10 (Stats SA, 2017: 218).

[16] *Il_taxable* is an income list which draws together several different types of reported income or expenditure that are listed in Table A2. *Il_taxable* is calculated as the sum of *yem, yse, ypr, ypp, yot and xishler,* minus *tscee_s.*
[17] In combination this equals the composite variable *ttb_s*.