International Tax and Public Finance (2019) 26:1018–1047
https://doi.org/10.1007/s10797-018-9529-9

The effect of top incomes on inequality in South Africa

Janina Hundenborn[1] · **Ingrid Woolard**[1,2,3,4] · **Jon Jellema**[4]

Published online: 27 June 2019

Abstract
South Africa exhibits extreme levels of income inequality and is ranked as one of the most unequal countries in the world. In order to measure these severe levels of inequality, it matters how we account for the different parts of the income distribution. Although the approach has gained international attention, there has not been any attempt at combining tax administration data with household survey data in order to account for incomes at all parts of the distribution, and especially from the top of the income distribution in South Africa. This paper uses a novel technique to identify the optimal method of combining tax administration with household survey data. Our results show the dramatic effects of accounting for reporting bias in household surveys by using tax administration data. When combining the two data sets, we find a significant decrease in overall inequality of taxable income in South Africa between 2011 and 2014, the 2 years under observation. Nonetheless, income inequality in South Africa remains high. For our analysis, we use two waves of the National Income Dynamics Study, a national representative household survey and compare the information to a sample of almost 1.2 million records on personal income tax for the 2011 tax year and about 1 million records the 2014 tax year.

Keywords Income distribution · Inequality · South Africa · Personal income tax

JEL Classification D31 · H24

✉ Janina Hundenborn
jhundenborn@gmail.com

1 Southern Africa Labour and Development Unit (SALDRU), University of Cape Town, Cape Town, South Africa

2 IZA, Bonn, Germany

3 World Institute for Development Economics Research (UNU-WIDER), Helsinki, Finland

4 Commitment to Equity Institute (CEQ), New Orleans, USA

1 Introduction

Acute levels of income inequality have led to South Africa being ranked as one of the most unequal countries in the world. Investigating the causes and consequences of such high levels of inequality has always been a key concern for scholars and policy makers. Pertaining to the analysis of income inequality, the effect of top incomes on overall inequality is gaining attention in the recent literature. How we account for this part of the income distribution matters since it has widely been acknowledged that household surveys tend to understate top incomes due to a higher non-response rate among high-income households or underreporting of incomes earned (see Atkinson et al. 2011; Van Der Weide et al. 2016). In this case, using tax income data has become a commonly used remedy to estimate incomes at the top end. Arguably, tax administration data are more precise on the top of the income distribution, however, they offer less detail on the bottom end of the distribution, particularly for individuals earning below the tax filing threshold (Morelli et al. 2014). Therefore, empirical evidence (Atkinson 2007; Burkhauser et al. 2012; Alvaredo and Londoño 2013; Diaz-Bazan 2015) has shown that tax data estimates on the top tail can be combined with estimates on the bottom segment of the population obtained from household survey data in order to estimate the entire income distribution. However, we will show that the existing literature that uses both, tax records and household survey data, to estimate the income distribution from top to bottom has not identified the optimal point at which the two types of data sets should be combined.

In the literature on income distributions, the analysis of the development of top incomes using tax data has gained substantial traction across countries. Garbinti et al. (2016), for example, study the evolution of top incomes using tax data from France, Atkinson (2007) uses tax data for the UK, Dell (2005) for Germany and Switzerland, Diaz-Bazan (2015) studies tax income combined with household survey data in Costa Rica and Alvaredo and Atkinson (2010) analyze top incomes in South Africa using tax data.

If external data such as tax administration data are combined with household survey data for the analysis of incomes, there are different techniques available to the researcher. As such, the analysis may combine household survey data for the bottom of the distribution with (more reliable) tax administration data on the top tail. This method was used by Garbinti et al. (2016) and Burkhauser et al. (2012), for example. Alternatively, inequality measures rather than data can be combined to assess the overall income distribution. This was done in the papers by Atkinson (2007) and Alvaredo and Londoño (2013) as well as Diaz-Bazan (2015), Lakner and Milanovic (2016) and Jenkins (2017). The methodologies of the latter will be discussed in more detail later on. Additionally, Lustig (forthcoming) offers a detailed review of alternative methods that correct for high earners in the absence of external data sources. These methods include the replacing of household survey data at the top end through semi-parametric estimates or imputations as well as the re-weighting (post-stratification) of incomes at the top as well as on the bottom of the income distribution. Lustig (forthcoming) highlights the fact that if the true distribution and the sample have the same support, any results obtained by replacing top earners can be replicated using the re-weighting method.

In the existing literature on South Africa, both, tax administration data and household survey data, have been used to assess the distribution of income. However, so far, these two sources of information have been studied separately in the South African context but not together. This paper will extend existing analysis by using a novel method to combine information from a unique data set on personal income tax with data from a South African household survey in order to study inequality across the entire distribution. We will show that the methods chosen are the best way to assess overall income inequality as close to the actual level of inequality as possible in the context of the large share of low-income earners and steep levels of inequality observed in South Africa.

The first set of data for the analysis of this paper stems from the National Income Dynamics Study (NIDS), a nationally representative household survey. Our analysis uses two waves of this study, Wave 2 from 2010/2011 and Wave 4 from 2014/2015. The attrition rates reported for the different income deciles in the NIDS panel show that the study is no exception to the common pattern of households at the top end showing increased non-response (see Finn et al. 2012 for a discussion on Wave 2, Table 7 in the Appendix for Wave 4). For this reason, we use personal income tax (PIT) data provided by the South African Revenue Service (SARS) to better estimate inequality for high-income earners. Individuals who earn above a certain income threshold are required to file taxes. Therefore, the information on incomes above this threshold is likely to be more accurate in the tax income data than the data from the household survey. The data sets provided on personal income tax comprise a 20% sample for the 2011 and the 2014 tax year each. However, having access to two data sets on individual income does not guarantee that the data sets are comparable. Hence, this paper reviews the information provided by the two data sources in great detail before applying the methodology introduced by Diaz-Bazan (2015) that combines the information from the two data sources to measure inequality across the entire distribution.

Our measurements of inequality show a decline in income inequality between 2011 and 2014. This decrease can be found across data sources and different methods of combining the two types of data. The fact that this decrease is even stronger when tax administration data and household survey data are combined using the method proposed by Diaz-Bazan (2015) highlights the need to account for distortions in household survey data by using tax administration data. Additionally, our discussion shows that for the broader part of the income distribution, household survey data reports relatively accurately on individual taxable income compared to tax administration data. However, we identify some weaknesses at the extreme ends of the distribution which makes the use of a combined analysis including tax administration data the preferred choice.

The remainder of this paper is structured as follows. Section 2 provides some macroeconomic context as well as an overview of the South African tax system before Sect. 3 introduces the two types of data used in our analysis. The distribution of taxable income in the different data sets is discussed in great detail in Sect. 4. The core of this paper builds the discussion of inequality in Sect. 5. Finally, Sect. 6 summarizes our findings and concludes.

2 Background

This section serves to give context and background information on the macro-economic and fiscal situation South Africa finds itself in 20 years after the end of apartheid. This is relevant to understanding the complexities that lead to continuously high levels of inequality despite broad efforts by the government towards redistribution and socio-economic inclusion.

2.1 Macro-economic context

The issue of inequality is particularly ripe in the South African context, yet the results of the government's commitment towards greater economic equality and redistribution in the post-apartheid era are coming up short for the majority of the population. When analyzing the effectiveness of fiscal policy in redistributing income, Inchauste et al. (2015) show that in South Africa the levels of poverty and inequality remain among the highest in middle-income countries. This is despite highly progressive social spending programs and the (slightly) progressive design of the overall tax system and the personal income tax in particular. What is most troubling is the fact that even though fiscal policy significantly reduces the Gini coefficient of income, it is still higher after this reduction than the Gini coefficient of other middle-income countries before policy intervention (Inchauste et al. 2015). Overall, South Africa does not fare well in comparison with other middle-income countries when it comes to inequality. Income per capita increased in South Africa between 1992 and 2012 as it did in other middle-income countries such as Brazil, Mexico or Thailand (Levy et al. 2015). Brazil, Mexico and Thailand are chosen due to the fact that they share similar characteristics with South Africa. As such, the average income, population size and therefore the development challenges faced by these countries can be assumed to be comparable (Levy et al. 2015). However, several studies find that the level of inequality in South Africa continues to exceed those of other middle-income countries. Leibbrandt and Finn (2012) find a Gini coefficient of household income per capita of 0.7 in South Africa in 2008, whereas Brazil reported a Gini coefficient of 0.55 in the same year. The stark levels of inequality translate into higher levels of poverty in South Africa compared to its peers. Although progressive fiscal policy achieves a significant reduction in poverty, the headcount ratio in South Africa measured at the US$ 2.50 per day poverty line remains high at 36% compared to Brazil, for example, at 11% (Inchauste et al. 2015).

Since the end of apartheid, the South African government has efficiently expanded fiscal programs and broadened the tax base in order to reduce poverty and inequality. However, these extensive efforts have not translated into the equivalent results. Levy et al. (2015) point out that "relative to other middle-income countries, South Africa has an unusually small fraction of the population that gains directly from sustained economic growth". Additionally, Table 1 shows that even though the annual growth rate increased up to 5.6% since the end of apartheid, growth slowed down significantly since the global financial crisis in 2007/2008. The decline of the growth rate limits the degree for possible expansion of progressive social spending. Coupled with the

Table 1 Real annual growth rates in South Africa since 1990. *Source*: Statistics South Africa (1993–2017)

Year	Growth rate (%)	Year	Growth rate (%)
1990–1993	− 0.5	2008–2010	1.8
1994–1999	2.8	2011–2015	2.1
2000–2004	3.8	2016	0.6
2005–2007	5.6	2017	1.3

previously high levels of fiscal debt and a large fiscal deficit, these macro-indicators allow little room for further fiscal policies to bring about greater redistribution. This problem may be aggravated since the downgrading of South Africa's debt to "junk status" by international rating agencies.[1] In order to significantly reduce poverty, unemployment and inequality, the National Development Plan[2] foresees an average annual growth rate of 5.4% until 2030. Table 1 highlighted the degree to which the growth rates fall short on this ambitious plan. It is for that reason that we will sharpen our focus on measuring income inequality after an overview of the South African tax system in the following section. The remainder of this paper will focus on the current levels of income inequality prevalent in South Africa and novel methods on how to optimize the assessment of inequality across available data.

2.2 The tax system in South Africa

South Africa's tax base is broad-based and generates a relatively high level of fiscal resources by middle-income country standards (Inchauste et al. 2015). Tax collections in 2016/2017 amounted to 26.2% of GDP, with a 60:40% split between direct and indirect taxes. Of the direct taxes, roughly two-thirds come from personal income tax. Personal income tax (PIT) is a tax levied on the taxable income (gross income less exemptions and allowable deductions) of a person. Capital gains also form part of taxable income. Individuals generally receive most of their income as salary/wages, pension/annuity payments and investment income (interest and dividends). Some individuals, such as sole proprietors and partners, may also have business income which is taxable as personal income.

The South African system of personal income tax is extremely simple. Filing is done individually, and the system does not distinguish between married and unmarried persons or provide deductions for children. There is, however, a small additional tax rebate for persons over the age of 65. All formal sector employees must be registered by their employer for PIT, and the employer is responsible for calculating and withholding the PIT payable by the employee. A certain level of interest income is tax-exempt in an effort to promote saving. Limited deductions are permitted for travel expenses and contributions to pension funds and medical aid (health insurance) schemes.

Over the past two decades, there has been considerable effort on the part of the tax authorities to broaden the tax base and to adapt the tax system to conform to

[1] Standard and Poor's rating agency rated South Africa's debt as speculative grade or "junk" in April 2017, other international rating agencies are currently reviewing the South African status.

[2] National Planning Commission (2011): "National Development Plan 2030, Our Future-make it work".

international tax laws. Fundamental changes included changing from a source-based to residence-based system in 2001 and the introduction of capital gains taxation to extend the tax base and enhance the equity of the tax system. There is a general consensus that the reforms to PIT made the system simpler and more equitable. The modernization of the tax administration system and reporting requirements imposed on financial institutions has resulted in extremely high levels of compliance. This makes the PIT data a particularly robust source of information about individual income.

3 Data sources

The analysis of income inequality in this paper is based on two main data sources. The first set of data stems from two waves of the National Income Dynamics Study. Secondly, the South African Revenue Service (SARS) provided a 20% sample of anonymized records on personal income tax records for the 2011 and 2014 tax years.

The analysis in this paper uses Wave 2 of the National Income Dynamics Study (NIDS) from 2010/2011 as well as Wave 4 from 2014/2015. NIDS is the first national representative panel survey in South Africa. The cross-sectional data sets used in this paper contain 34,000 individuals for 2010/2011 and about 42,000 individuals for 2014/2015. The large scope of this study is one of the main advantages of using the NIDS survey data. Furthermore, NIDS contains detailed information on incomes from primary and secondary employment, self-employment and a list of bonuses that is used to asses labour income. Information on rent and interest earned is used to estimate investment incomes comparable to the information in the tax data. It is important to note that there is an exemption for local interest earned of up to R23,800[3] for individuals below 65 years and up to R34,500[4] for individuals 65 years and older. Due to the nature of the questionnaire, it is not possible to distinguish between local and international interest earned in NIDS. This may inflate taxable income in NIDS slightly; however, as only a small fraction of individuals report investment income, the effect should be negligible. Capital incomes in NIDS can be estimated using the available information on retrenchment payments, loans, sales of household assets and other income of a capital nature. With these components, it is possible to construct an income variable in NIDS that is comparable to the information available in the tax data. A detailed comparison of the data provided by SARS and the information available in the NIDS data are included in Table 9 in the Appendix. In order to further ensure that income variables in the two data sources are corresponding, all values of the 2014/2015 NIDS survey as well as the 2014 PIT data have been deflated to prices that reflect the 2011 tax year.

The South African tax year spans from March of the previous year to February of the current year. PIT data on the 2011 tax year therefore span from March 2010 to February 2011. The data on personal income tax (PIT) provided by SARS contain almost 1.2 million records for the 2011 tax year and about 1 million records for the 2014 tax year. The PIT data contain detailed information on investment incomes

[3] In PPP$, R23,800 in 2014 is equivalent to $4985 in 2011 prices (OECD 2017).

[4] In PPP$, R34,500 in 2011 is equivalent to $6689 (OECD 2017).

including capital gains, business profits and losses. Furthermore, it provides data on labour income including income taxed as "pay as you earn" (PAYE) and fringe benefits as well as lump-sum incomes linked to earnings. Additionally, the PIT data include any possible exemptions and deductions as well as some demographic information such as age, gender and office of registration. The information on incomes is summed up in a variable on taxable income which forms the base for the tax payer assessment by SARS. The data provided by SARS are a 20% sample of all individuals who filed their tax returns. It is important to note that this includes individuals who filed tax reports despite earning incomes below the mandatory tax filing threshold. Tax filing is mandatory for individuals who earn above a certain threshold, for the 2011 tax year, that threshold was R120,000[5] annual income. In 2014, however, the threshold had increased to R250,000[6] annual income. When we correct this threshold for inflation, R250,000 is equivalent to R199,397 in 2011 prices.[7]

The following section will provide a more detailed analysis of the distribution of the taxable income variable in the household survey data and in the tax administration data.

4 Distributional analysis

4.1 Overview of the household survey data

The previous section discussed the components of the variable of taxable income created in the household survey. In this section, we are going to review the distribution of taxable income in more detail.

Table 2 offers a brief overview of the distribution of taxable income in NIDS for adults 18 years and older. Distributional statistics such as the first, 25th, 50th, 75th, 90th, 95th and 99th percentile are reported for both years. Additionally, the table shows the first nonzero percentiles for both years. In 2011, more than half the population reports zero taxable income. Only at the 53.8th percentile, can we observe positive taxable income. In 2014, less than half the population earns zero taxable income as the first positive incomes are observed at the 46.3rd percentile. Furthermore, Table 2 highlights the percentiles that are closest to the different filing thresholds of the two years. In 2011, the tax filing threshold was at R120,000[8] which is equivalent to the 93.6th percentile in the 2011 household survey data. The filing threshold in 2014 was at R250,000[9], accounting for inflation, that is equivalent to R199,397 which lies between the 97.4th and the 97.5th percentile of the 2014 NIDS distribution. If the 2014 threshold were to be applied to the 2011 NIDS distribution, the inflation-adjusted threshold would lie between the 97.6th and the 97.7th percentile. In the existing liter-

[5] In PPP$, R120,000 is equivalent to $25,136 (OECD 2017).

[6] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

[7] For simplification, we will continue to refer to it as the R250,000 threshold although all prices have been deflated to 2011 levels for the remainder of this paper.

[8] In PPP$, R120,000 is equivalent to $25,136 (OECD 2017).

[9] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

Table 2 Taxable income—percentiles in the NIDS data. *Source*: Authors' calculations using NIDS (weighted)

Percentile	2011	2014
1%	0	0
25%	0	0
46.3%		453
50%	0	7654
53.8%	300	
75%	23, 700	29, 457
90%	82, 548	84, 449
93.6%	120, 000	
95%	148, 200	140, 030
97.4%		196, 464
97.5%		200, 383
97.6%	197, 200	
97.7%	202, 500	
99%	324, 000	413, 154
Mean	30, 220	36, 377

ature on combining tax data with household survey data, different percentiles are used as connection points and we will argue later in this paper which one we believe is best in the context of the South African income distribution.

The distribution of taxable income shown in Table 2 highlights one of the major problems when it comes to inequality in South Africa. In 2011, more than 50% of individuals have zero taxable income. There is an apparent upward shift in the distribution in 2014 which can be seen as only about 46% report zero taxable income. However, it is very important to note that the individuals who report no taxable income may receive income from other sources. Taxable income is comprised of labour income, business profits, capital income and other sources liable for taxation. The large share of individuals that earn no taxable income highlights the strong dependence on other income sources for individuals at the bottom of the income distribution. Non-taxable income sources include government grants which specifically target poor individuals; Schiel et al. (2016) discuss the strong dependency on government grants in more detail. Other non-taxable income sources include intra-household transfers where household members profit from another household member's income. This type of resource sharing cannot be accounted for when analyzing taxable income which is only available at the level of the individual. Finally, inter-household transfers in the form of remittances may be another source of income that cannot be taxed and therefore will be overlooked in our analysis.

The aforementioned upward shift in the 2014 income distribution can also be seen by the development of mean incomes reported in the bottom row of Table 2. Mean incomes increased from R30,220[10] in 2011 to R36,377[11] in 2014 even though all other

[10] In PPP$, R30,220 is equivalent to $6330 (OECD 2017).

[11] In PPP$, R36,377 is equivalent to $7620 (OECD 2017).

percentiles in 2014 are less than their 2011 equivalents, except for the 99th percentile. This is another indicator of the high levels of income inequality prevalent in South Africa. In 2011, the mean taxable income for individuals was at R30 220, whereas the median (50th percentile) was at zero. This is a strong indicator that the top of the income distribution is moving away from the median, implying widening levels of inequality. At a median income level of R7654[12] and a mean of R36,377 in 2014, these stark levels of inequality do not seem to have subsided. The analysis of mean and median points of a distribution is a reliable indicator of inequality,[13] but the substantial differences between the top of the income distribution and the bottom warrant further investigation of these income groups. The following section will review higher income earners in the tax data more closely.

4.2 Overview of data on taxpayers

The PIT data contain its own taxable income variable, which in the following sections is shown in raw (non-manipulated) form. We use this data provided to create a proxy for mean taxable income by income bracket. This proxy is created based on the continuous distribution of taxable income in the PIT data. Table 3 below lists 24 income brackets including zero incomes in the PIT data and the percentage of total taxpayers in each bracket.[14] Even though the data on personal income tax provided are continuous, these results are reported in brackets in order to relate this discussion to the annual publication of the Tax Statistics by SARS (2015).[15]

When looking at the two years for which SARS provided data on taxpayers, there is a clear shift towards the upper end of the distribution between 2011 and 2014. In the PIT data, there are relatively more taxpayers in each income bracket above a yearly income of R130,000 in 2011 prices. At the same time, the number of tax filers reporting zero income reduced to half of its 2011 level, from 4.53 to 2.34% in 2014. It is important to view any statistics below the filing threshold with a certain level of caution as only a few individuals with lower incomes will have filed for taxes. Many income earners will not be captured due to voluntary self-reporting of incomes at this level. Of those who have filed taxes, many will have done so in order to gain refunds. As mentioned above, filing for taxes is compulsory only once an individual earns above the filing threshold in the specific tax year. To recall, the filing thresholds were at R120,000[16] for 2011 and at R250 000[17] in 2014. The inflation-adjusted threshold in 2014 is at R199,397. We will show later on in the paper that only above these filing thresholds are the PIT data more reliable.

[12] In PPP$, R7654 is equivalent to $1603 (OECD 2017).

[13] See Wittenberg (2016) for a discussion on mean and median in the context of wage inequality in South Africa.

[14] Table 10 in the Appendix reports the same income brackets in PPP$.

[15] The 2015 Tax Statistics are cited here as it is the last year that reports for both the 2011 and 2014 tax year.

[16] In PPP$, R120,000 is equivalent to $25,136 (OECD 2017).

[17] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

Table 3 Income brackets in the PIT data. *Source*: Authors' calculations based on PIT Data (2011 and 2014)

Income group	2011		2014	
	Mean taxable income	Percentage of taxpayers (%)	Mean taxable income	Percentage of taxpayers (%)
0	0	4.53	0	2.34
1–20,000	9630	4.97	10, 083	3.77
20,000–30,000	25, 130	2.58	25, 088	2.03
30,000–40,000	35, 146	2.94	35, 136	2.27
40,000–50,000	45, 211	3.34	45, 273	2.81
50,000–60,000	55, 419	4.59	55, 024	3.81
60,000–70,000	65, 109	4.82	65, 048	4.02
70,000–80,000	74, 925	5.30	75, 083	4.36
80,000–90,000	85, 000	4.63	84, 971	4.39
90,000–100,000	94, 967	4.04	94, 982	4.24
100,000–110,000	104, 889	3.84	104, 946	4.00
110,000–120,000	115, 132	3.67	114, 987	3.93
120,000–130,000	124, 965	3.66	124, 929	3.59
130,000–140,000	134, 934	3.32	135, 108	3.61
140,000–150,000	145, 012	3.09	144, 824	3.68
150,000–200,000	173, 340	13.73	174, 203	14.63
200,000–300,000	241, 192	13.12	241, 559	16.22
300,000–400,000	344, 656	5.63	344, 453	6.70
400,000–500,000	444, 726	2.91	444, 974	3.47
500,000–750,000	600, 010	3.11	598, 698	3.64
750,000–1,000,000	855, 902	1.01	853, 815	1.16
1,000,000–2,000,000	1, 322, 742	0.90	1, 325, 424	1.02
2,000,000–5,000,000	2, 860, 366	0.24	2, 873, 280	0.27
Above 5,000,000	7, 704, 813	0.03	7, 571, 683	0.03
N_weighted	5, 783, 360	100	5, 085, 060	100

There is a significant jump in mean taxable incomes that can be observed for the highest income bracket including individuals that earn R5 million a year and more.[18] SARS provided additional statistics on high earners with annual taxable incomes of R10 million or more.[19] These individuals are included in this bracket and drive up the mean.

4.2.1 High earners in the PIT data

The South African tax authority has provided additional statistics for top income earners who report taxable incomes above R10 million a year for each 2011 and 2014. In order to protect their identity, SARS has only provided summary statistics for earners

[18] In PPP$, R5,000,000 is equivalent to $1,047,340 in 2011 prices (OECD 2017).

[19] In PPP$, R10,000,000 is equivalent to $2,094,680 in 2011 prices (OECD 2017).

Table 4 High earners' income distribution. *Source*: Authors' calculations based on PIT Data (2011 and 2014)

Quartiles	2011 Income	2014 Income
Min	10,005,439	7,978,814
Q1	11,551,323	9,302,885
Q2	14,078,782	11,774,718
Mean	16,987,843	15,783,480
Q3	17,844,600	16,861,538
Max	109,909,354	151,159,382
Number of Obs	482	1048

above this threshold, including the number of earners, maximum and minimum as well as mean and different quartiles of the distribution of top income earners. In 2011, there were a total of 482 individuals who reported a taxable income above R10 million, and in 2014, this number had increased to 1048 individuals. SARS took these samples without accounting for inflation. In order to make the data of the two years comparable, prices have been adjusted for inflation in our analysis. Therefore, the 2014 levels appear much lower in comparison with the different statistics provided by SARS between 2011 and 2014 in Table 4. In 2011, the smallest income reported in this segment of the population was at R10,005,439 and the first quartile can be observed at R11.5 million. The smallest value observed in the 2014 data on high earners was at R7,978,814 in 2011 prices. In real terms, the upward shift in the distribution mentioned in the discussion of Table 3 cannot be observed here except for the highest value reported. It is important to note that the median or second quartile is lower than the mean or average. The median in 2011 was at R14 million and in 2014 at R11.7 million. The fact that the mean is significantly higher in both years with a value of close to R17 million in 2011 and R15.8 million in 2014 indicates that the maximum values above R100 million in both years are outliers that drive up the average, whereas half of these high-income earners actually report R14 million or less annual income.

The information on taxable income of these high-income earners discussed in this section will be included in the PIT data for the forthcoming analysis of inequality across different data sources.

4.3 Comparison of household survey data with tax administration data

As discussed in Sect. 3 above, we derive a taxable income variable in NIDS for individuals 18 years and older. In order to reconcile our analysis with the annual Tax Statistics published by SARS (2015), Sect. 4.2 then discussed the PIT data with regards to 24 income brackets with taxable income limits greater than zero. Figure 1 assesses the amount by which the taxable income variable in NIDS is greater or smaller than the PIT taxable income. For this purpose, the mean of each of the 24 taxable income brackets has been calculated in NIDS and in PIT. The results show that, with the exception of the extreme ends of the distribution, the difference of these mean taxable incomes across the different income brackets is negligible between NIDS and PIT.[20]

[20] Figure 6 in the Appendix shows Fig. 1 but with income brackets in PPP$.

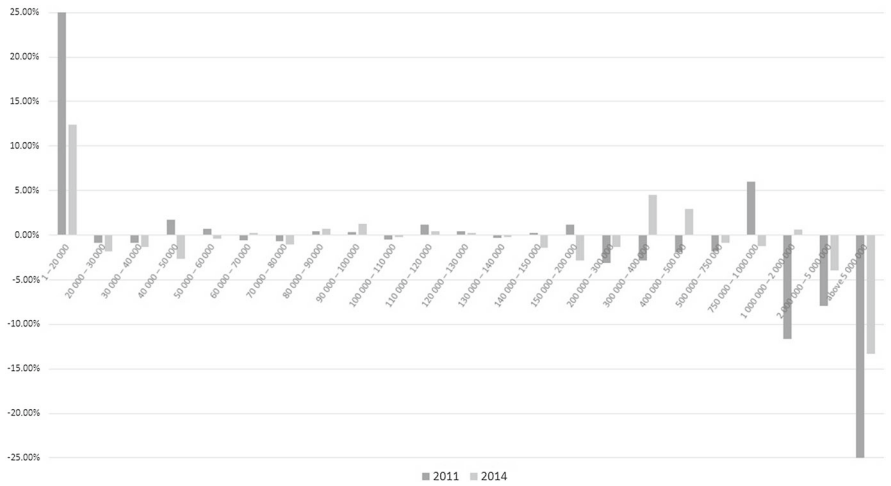

Fig. 1 Difference in estimated mean taxable income between NIDS and PIT by income bracket

At the bottom end of the distribution, NIDS estimates low-income earners at much higher percentages when compared to the PIT data. In 2011, the difference is as large as 25% for positive incomes below R20,000. However, the difference of mean taxable income between NIDS and PIT decreased to 12% for the lowest income bracket above zero in 2014. This difference at the bottom of the distribution is hardly surprising as not many individuals will have filed taxes when earning such small incomes and are therefore missing in the tax administration data but will have been covered by the household survey.

In 2011, any individual earning R120,000[21] or more had to file tax returns. Around this threshold, the taxable income proxy created in NIDS reports very close estimates to the mean taxable incomes reported in the PIT data for both 2011 and 2014. The estimated discrepancies between the two data sources start increasing for incomes of R200,000 and higher. In Table 2, we have shown that R200,000 is at the very top of the income distribution in the household survey in 2011 and in 2014. At this point of the distribution, NIDS understates mean taxable income by about 3% in 2011 but only by about 1% in 2014. In 2011, the differences between the two data sets remain small until the income brackets above R750,000. For incomes between R750,000 and R1,000,000, NIDS actually overestimates mean taxable income by about 6%. For the remaining higher income brackets, NIDS severely underestimates mean taxable income in 2011 with a difference of up to 25% compared to PIT for incomes above R5,000,000.

The mandatory filing threshold increased to R250,000[22] in 2014. For incomes between R200,000 and R300,000, NIDS underestimates mean taxable income slightly. For the income brackets between R300,000 and R500,000, however, NIDS actually overestimates mean taxable income by between 3% and 5%. These are surprising

[21] In PPP$, R120,000 is equivalent to $25,136 (OECD 2017).

[22] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

findings as the previous literature would indicate that households at the top end of the income distribution tend to under-report their incomes and not over-report as this would suggest. A possible explanation to this conundrum lies in the high attrition rates of top income households which will have affected the fourth wave of NIDS in 2014 to a larger degree than the second wave in 2011.[23] Due to increasingly high non-response rates of high-income households, those few that remain in the study will be attributed larger weights to ensure continued national representation of the study. It is possible that this has led to a bias through over-weighting this top end of the income distribution. We will further investigate this issue in the following sections.

Furthermore, NIDS underestimates the incomes in the very top income brackets in 2014, however, to a much smaller degree than in 2011. For incomes between R2,000,000 and R5,000,000, NIDS estimates 4% less mean taxable income than PIT. For incomes above R5,000,000, the difference is just above 12%, about half of the difference in 2011. For all other income brackets, the differences between the proxy taxable income variable in NIDS and taxable income provided in the PIT data see-saw around zero. Therefore, the under- and over-estimation of mean taxable incomes appear to behave somewhat like random fluctuation around the true mean in both years. In other words, NIDS appears to perform reasonably in capturing mean taxable incomes from respondents.

However, while taxable income levels and overall magnitudes in NIDS correspond reasonably well to the same levels contained in the PIT data, inequality is determined not by the overall magnitude of the cumulative total but its distribution. PIT and NIDS demonstrate substantial differences in the distribution of total taxable income, especially at the extreme ends of the distributions. As these ends have a significant impact on the overall measurement of inequality, the next sections will pursue these differences further in order to uncover what impact the data source has on the overall level of inequality in both gross and net incomes.

5 Measuring inequality

Traditionally, inequality in South Africa has been studied using household survey data (see for example Leibbrandt et al. 2012) and labour force survey data (see for example Wittenberg 2016). Alvaredo and Atkinson (2010) provide an exceptional analysis of top income shares over a one hundred-year period using South African tax data but are unable to utilize it to assess inequality. To the authors' knowledge, no study has attempted to use a combination of household survey data and tax administration data to assess income inequality across the entire income distribution in South Africa. As a consequence, we are able to contribute to the existing literature on income inequality through access to a unique set of tax administration data and high-quality household survey data on South Africa.

[23] See Finn et al. (2012) for a discussion of attrition rates in Wave 2 and Table 7 in the Appendix for Wave 4.

The previous sections have discussed the two data sources used for this analysis in great detail. The discussion of the distribution in the household survey data has highlighted a large number of individuals earning zero taxable income. However, the number of these zero earners has decreased from 2011 to 2014 and mean incomes have increased. The vast discrepancy between the mean and median observed in the NIDS data indicated relatively high levels of inequality without using statistically more advanced methods. As in the household survey data, tax administration data indicated an upward shift in the overall distribution of taxable income. The discussion of extremely high earners in the PIT data showed a discrepancy between mean and median incomes, highlighting extreme outliers that drive up the average. This section is using more advanced methods to look at inequality in the two data sets separately and how best to combine these two data sets to improve on existing analysis of income inequality.

At this point, it is important to recall that we measure inequality of taxable income of individuals rather than households. These measurements are going to be higher than inequality measures at the household level for two reasons. Firstly, our analysis measures taxable income which means other income sources relevant particularly to individuals at the bottom of the income distribution such as remittances and government grants will be left out. Several studies have shown the immense impact that government grants had on lowering income inequality over the past years (Inchauste et al. 2015; Hundenborn et al. 2016) so the large fraction of zero earners will have a significant effect on our measures of inequality. Secondly, in a household, there is sharing of resources such that even if one household member earns no income, the household as a whole may still report some form of revenue through another person's income. Taxable income of the individual by definition cannot account for such sharing of resources. The measurements of inequality of individual taxable income will be discussed in more detail in the following section.

5.1 Inequality within the NIDS and PIT Data

A commonly used measurement for inequality is the Gini coefficient. Table 5 presents the Gini coefficient for the household survey data for the different years. We have argued previously that the tax administration data are less reliable below the tax filing threshold as fewer individuals will have filed taxes below that limit. Therefore, the analysis of the Gini coefficient is broken up into segments above and below the tax filing threshold.

In 2011, the mandatory tax filing threshold was at R120,000.[24] Looking at taxable income in the NIDS data, the overall Gini coefficient in 2011 was relatively high at 0.823. Below the R120,000 filing threshold, inequality is measured at 0.762. As the top end of the distribution is truncated, the level of inequality decreases. Among the top income earners, we observe much lower levels of inequality at 0.359. In 2014, a noticeable decrease in inequality can be examined. Table 2 indicated a significant reduction in the zero earners paired with an overall upward shift of incomes earned which might explain the decrease to a Gini coefficient of 0.813 in 2014. The mandatory

[24] In PPP$, R120,000 is equivalent to $25,136 (OECD 2017).

Table 5 Gini coefficients at different thresholds. *Source*: Authors' calculations using NIDS and PIT (weighted)

Threshold	2011		2011 at 2014		2014	
Data source	NIDS	PIT	NIDS	PIT	NIDS	PIT
Overall	0.823	–	0.823	–	0.813	–
Below filing threshold	0.762	–	0.783	–	0.735	–
Above filing threshold	0.359	0.367	0.369	0.326	0.472	0.349

tax filing threshold in 2014 was at R250,000.[25] Despite this considerable increase in the filing threshold, the level of inequality below this threshold is significantly lower at 0.735 compared to the 2011 values. At the same time, the level of inequality above the filing threshold has increased substantially from 0.359 in 2011 to 0.472 in 2014.

The spike in inequality above the filing threshold as measured by NIDS in Table 5 is somewhat concerning. Therefore, Table 5 also compares the Gini coefficients above the respective tax filing threshold for 2011 and 2014 to the PIT data sets. The estimation of the Gini below the filing threshold would be biased in the PIT data and is therefore not reported. The large increase in inequality above the filing threshold in the household survey data cannot be observed in the tax administration data for 2014. In fact, according to the PIT data, inequality above the filing threshold decreased from a Gini coefficient of 0.367 in 2011 to a Gini coefficient of 0.349 in 2014.

Due to the significant shift in filing thresholds, one cannot simply compare the levels of inequality in 2011 with inequality in 2014. For this reason, Table 5 assesses the level of inequality below the 2014 filing threshold in the 2011 data. As the filing threshold increases, the level of inequality below the threshold increased from 0.762 to 0.783 in the NIDS data. At the same time, the 2011 NIDS data also report a higher Gini coefficient above the 2014 threshold. At this level, the Gini coefficient increased from 0.359 to 0.369. However, the PIT data report a decrease in the Gini coefficient from 0.367 to 0.326 when the filing threshold is shifted from R120,000 in 2011[26] to R250,000 in 2014.[27] Comparing the level of inequality prevalent in 2011 with 2014 once the same filing threshold is applied, we observe a decrease in inequality below the filing threshold in the NIDS data, while inequality above the filing thresholds increased in both data sets.

As mentioned in our discussion of Fig. 1, it can be assumed that the PIT data are much more precise in reporting higher incomes than the household survey data. Therefore, the discrepancies observed in Table 5 indicate a potential weighting issue in the NIDS data set in 2014. Such weighting issues can be caused by higher refusal rates in the upper income deciles compared to lower income deciles (see Table 7 in the Appendix) or outliers in the household survey data. It is for these reasons that tax administration data are the preferred source of information to assess income inequality at the top of the income distribution.

[25] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

[26] In PPP$, R120,000 is equivalent to $25,136 (OECD 2017).

[27] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

5.2 Inequality across the NIDS and PIT data

The comparison of the Gini coefficients in Sect. 5.1 provides some insight into the overall level of inequality as well as the difference in inequality measured in the NIDS data and the PIT data. Both of these types of data sets suffer from certain shortcomings. Previous studies have argued that household survey data suffer from distortions at the top end of the distribution (for example Van Der Weide et al. 2016 or Atkinson et al. 2011). Tax administration data, on the other hand, are much less reliable below the filing threshold as we do not know who the tax filers are that report their incomes at this point of the distribution. Therefore, the tax records below the filing threshold will not be representative. In order to understand the distributions of the household survey data and tax administration data better, Figs. 2, 3, 4 and 5 plot the kernel density functions of the NIDS vis-á-vis the PIT data. The results draw a very clear picture of the shortcomings presented above.

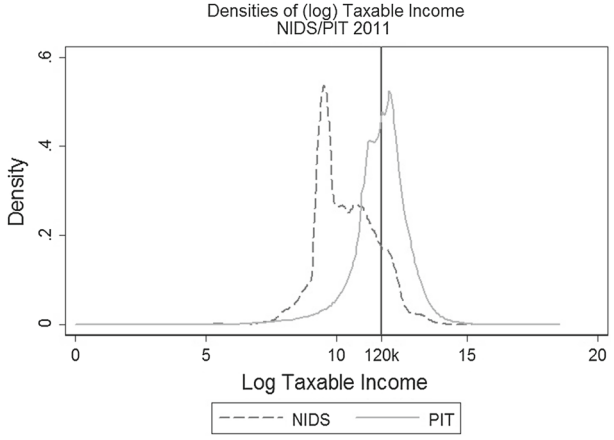

Fig. 2 2011 Taxable income in NIDS versus PIT

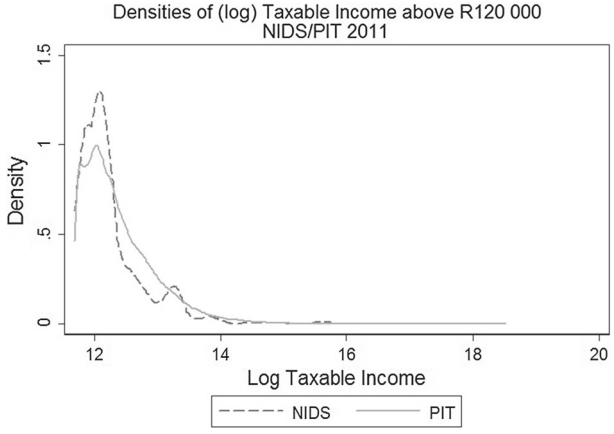

Fig. 3 2011 Taxable income in NIDS versus PIT above the filing threshold

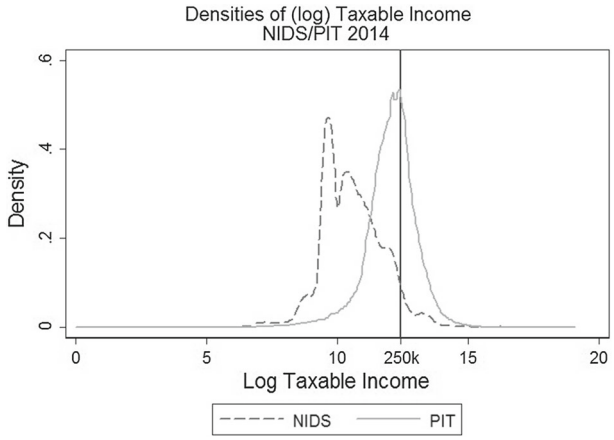

Fig. 4 2014 Taxable income in NIDS versus PIT

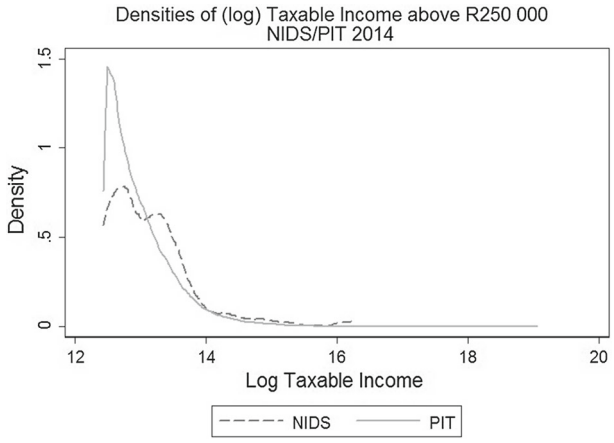

Fig. 5 2014 Taxable income in NIDS versus PIT above the filing threshold

Figure 2 reports the densities of taxable income in logs in the household survey data and the tax administration data. In this graph, the household survey seems to report many more incomes below the filing threshold of R120,000 but is outperformed by the tax data in capturing incomes above the filing threshold which is marked as *120k*.[28] NIDS thereby shows a spike in incomes below the filing threshold that PIT lacks completely, implying that there are a lot of incomes that were not reported in the PIT data but were successfully captured in the household survey. Due to the nature of the logarithmic transformation, Fig. 2 does not picture the numerous zero earners discussed previously, another relevant part of the distribution missed by PIT. On the other hand, the tax administration data start picking up incomes from just below the mandatory filing threshold that cannot be found in NIDS as evident by a spike in

[28] The logarithmic of R120,000 is about 11.7.

incomes around the *120k* mark that PIT captures but NIDS does not. Figure 3 takes a closer look at the distribution above this threshold.

The high spike around a log of 12, an actual income of about R162,754,[29] may be the reason NIDS performs reasonably well when comparing mean incomes by income bracket at these levels of the income distribution. However, the household survey data trail off at the very high income levels and the PIT data report incomes much higher than those reported in the NIDS data. The significance of the income levels above the filing threshold will be discussed in more detail in the following section.

Figure 4 shows a similar overall pattern for 2014. NIDS captures a lot more incomes below the filing threshold of R250,000 which is marked as *250k*.[30] Even though zero earners are again not pictured, NIDS captures a number of incomes very close to zero. The overall distribution in 2014 is rather similar to the 2011 distribution, and so Fig. 4 reports a large spike in incomes below the filing threshold as well. Again, the tax administration data fail to report any significant incomes at this segment of the distribution. However, as the graph moves closer to the mandatory filing threshold, PIT starts capturing incomes that NIDS does not report. Above the filing threshold, NIDS fails to capture a significant number of incomes reported in the PIT data. This is emphasized by the distribution pictured in Fig. 5.

In the graphs reported in Fig. 5, the tax administration data show a spike very close to the mandatory filing threshold that is missing in the NIDS data. Further up the income distribution, the NIDS graph moves to the right of the PIT distribution. This is a surprising observation as the existing literature focuses on the issue of underreporting of top incomes in household surveys. However, it would seem that the issue of missing observations in the top tail due to higher non-response of high-income households is more prevalent in the South African household survey data. The few high-earning individuals who are observed will receive larger weights in order to represent everyone in the top tail of the income distribution. It appears that after four waves of the household survey, the higher attrition rates among high earners have introduced a bias at this part of the income distribution which leads to this surprising over-reporting of taxable income in NIDS. Nonetheless, after a significant hump at a log of about 16, a real income of about R8.9 million,[31] the NIDS data fail to capture any further incomes. This outlier and the generally higher levels of income reported by NIDS up to that point may explain why inequality in the NIDS data was so much higher than in the PIT data as reported in Table 5; the distribution shown in Fig. 5 is a lot more equal in the PIT data than in the household survey despite the fact that PIT captures high earners (with a log income above 16) that are absent from the NIDS data.

The figures discussed in this section show that in both years, household survey data capture incomes at the bottom of the income distribution that are not captured by the PIT data. At the same time, the tax administration data start picking up incomes that are not reported in the NIDS data from income levels just below the respective filing thresholds. Additionally, Fig. 5 detected a potential weighting issue in the NIDS data above the filing threshold. Because of these characteristics, the following section

[29] In PPP$, R162,754 is equivalent to $34,092 (OECD 2017).

[30] Corrected for inflation, the logarithmic of the R250,000 threshold is at 12.2.

[31] In PPP$, R8,900,000 is equivalent to $1,864,265 in 2011 prices (OECD 2017).

will discuss a method of how to combine the two types of data in order to capture all segments of the income distribution and estimate the true level of income inequality more accurately.

5.3 Combining household survey data and tax administration data

The discussion of the two types of data sets in this paper has shown that the NIDS household survey captures the broad spectrum of incomes relatively well when compared to the tax administration data provided by SARS. However, the above figures highlight certain shortcomings in the capturing of the top incomes in the NIDS data. Since the top end of the income distribution is crucial to the analysis of overall inequality, it follows that tax administration data should be utilized to improve on the overall analysis of inequality. Previous studies on the topic support this idea.

Atkinson (2007) highlighted that changes in the incomes of the top income earners can significantly affect overall inequality. This holds true even when the group of top income earners may be rather small: "If we treat the very top group as infinitesimal in numbers, but with a finite share S* of total income, then the Gini coefficient can be approximated by S*+ (1 − S*) G, where G is the Gini coefficient for the rest of the population" (Pg. 19). In a numeric example, an increase of 8% in S* would then lead to an increase in the Gini coefficient by 4.8% if the Gini for the rest of the population (i.e. 1 − S*) was at 0.4 prior to the increase. Alvaredo (2011) further extends Atkinson's argument and asserts that depending on the degree of underreporting in household surveys, one can use tax data from the top 1% up to top 5% or 10%. However, neither Atkinson (2007) nor Alvaredo (2011) provide a guideline as to which threshold would be preferable in order to analyze the entire income distribution.

Because of this arbitrary choice of thresholds, there exist many studies that attempt to solve the problem of including high-income earners in the assessment of inequality in different ways. Jenkins (2017), for example, uses a more principled approach and determines the appropriate threshold through statistical testing. Using both Pareto Type-I and Type-II distributions to estimate the top end of the income distribution, he concludes that the Pareto Type-II estimates are more robust and that the threshold at which household survey data and tax administration data should be combined varies across years. Jenkins' (2017) detailed that a statistical approach is possible because of the availability of household survey data and tax administration data for the UK. Lakner and Milanovic (2016) use a different semi-parametric approach to combine inequality indices in order to account for high-income earners. Because of a lack of reliable tax administration data particularly in developing countries, the authors allocate the excess of national accounts private consumption over the mean of the household survey data to the top 10% for each country, thereby "elongating" the distribution using Pareto imputations. While this innovative approach is useful in the absence of reliable tax administration data, not only is the 10% threshold chosen rather arbitrarily but as Jorda and Niño-Zarazúa (2016) point out, the choice of Pareto imputations might not be the optimal distribution to use.

Alternatively to the semi-parametric approaches of Jenkins (2017) and Lakner and Milanovic (2016), Jorda and Niño-Zarazúa (2016) use a fully parametric approach to

estimate the threshold at which household survey data are truncated. For this, they argue that a general class of distributions, the so-called generalized Beta function of the second kind (GB2), is the most useful model specification as it has been shown to be an adequate fit to income data. The optimal truncation point is chosen to be that threshold at which the squared differences between the estimated top incomes using the GB2 distribution and the top incomes available from the tax administration data are minimized. This allows the optimal threshold to vary across countries. However, this is only possible for countries with tax administration data available. To account for high-income earners when measuring inequality in countries around the world, Jorda and Niño-Zarazúa (2016) have to resort to varying the threshold more intuitively and use GPD per capita as a proxy when tax administration data are not available. Therefore, while Jorda and Niño-Zarazúa (2016) use the optimal threshold where possible, the overall choice at which to combine household survey data with data that accounts for high earners remains arbitrary.

In an effort to actively account for distortions due to reporting bias and higher non-response rates of high earners in household survey data, Diaz-Bazan (2015) argues that rather than choosing an arbitrary percentile at which to combine the information from the two sources of data, the data sets should be combined at an optimal threshold b. This optimal threshold should fulfill the following conditions: (a) individuals with incomes below threshold b are well represented in the household survey data; (b) the segment of the population reporting incomes above b is captured accurately in the tax records; and (c) the threshold is chosen such that distortions attributed to underreporting at the top of the income distribution in household survey data are minimized. The optimal b that fulfills these conditions minimizes dependence on household survey data and ensures reliable information in the tax administration data. Therefore, the optimal threshold b is the lowest level of income that requires individuals to file taxes. As previously discussed, the mandatory tax filing in South Africa started at annual incomes of R120,000 in 2011[32] and R250 000 for 2014.[33] Following Diaz-Bazan's argument, these are the optimal thresholds for combining the two data sources. Table 2 above has shown that in South Africa, the mandatory filing thresholds are well below the 1% thresholds traditionally used in the existing literature of combining tax administration data with household survey data. To recall, the filing threshold in 2011 was at the 93.6th percentile and in 2014, the filing threshold was close to the 97.5th percentile. The following sections will discuss the methodology of the approach introduced by Diaz-Bazan (2015) as well as the results of its application to the NIDS and PIT data in 2011 and 2014.

5.3.1 Methodology following Diaz-Bazan (2015)

If the optimal threshold b that satisfies the conditions outlined above is at the filing threshold, the Gini coefficient for the entire (unconditional) income distribution can be estimated by combining the (conditional) Gini coefficient below b, where F_1 is estimated using the household survey data, with the (conditional) Gini coefficient above b where F_2 is estimated from the tax administration data such that

[32] In PPP$, R120,000 is equivalent to $25,136 (OECD 2017).

[33] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

$$F_1 = \Pr\{Y \leq y | Y \leq b\}$$
$$F_2 = \Pr\{Y \leq y | Y > b\}.$$

Given the definition of the conditional distribution, the distribution for $y \leq b$ is represented by

$$F_1(y) \equiv \Pr\{Y \leq y | Y \leq b\} = \frac{\Pr\{Y \leq y\}}{\Pr\{Y \leq b\}} = \frac{F(y)}{F(b)}$$

so that $F(y) = F_1(y)F(b)$.

Similarly, the distribution for $y > b$ can be written as

$$F_2(y) \equiv \Pr\{Y \leq y | Y > b\} = \frac{\Pr\{Y \leq y \text{ and } Y > b\}}{1 - \Pr\{Y \leq b\}} = \frac{F(y) - F(b)}{1 - F(b)}$$

so that $F(y) = F(b) + (1 - F(b))F_2(y)$.

This implies that it is possible to reconstruct the underlying income distribution $F(y)$ if the two conditional distributions above and below the threshold b can be observed and the fraction of the population with income above and below threshold b are known.

In order to obtain a measurement of inequality for the entire population, it is possible to estimate the Gini coefficient using a linear combination of the Gini coefficients computed on the conditional distributions above and below threshold b. In this case, the Gini coefficient of a conditional distribution, $G[F_j]$, $j \in 1; 2$ can be written as

$$G\left[F_j\right] = \frac{1}{\mu_j} \int_0^\infty F_j(y)(1 - F_j(y))\mathrm{d}y,$$

where $\mu_1 \equiv E[Y|Y \leq b]$ and $\mu_2 \equiv E[Y|Y > b]$. This G is equivalent to the standard Gini coefficient calculated on the two distributions.

This can then be combined into a full population, unconditional Gini coefficient defined as

$$G\left[F\right] = (1 - F(b))F(b)[\frac{\mu_2 - \mu_1}{\mu}] + \left(F(b)^2 \frac{\mu_1}{\mu}\right) G[F_1] + \left(F(b)^2 \frac{\mu_2}{\mu}\right) G[F_2],$$

where $\mu = F(b)\mu_1 + (1 - F(b))\mu_2$ and $G[F_i]\, i \in \{1, 2\}$ are the Gini coefficients of F_1 and F_2.

Therefore, the (unconditional) Gini coefficient can be written as a linear combination of the two Gini coefficients of the conditional distributions F_1 and F_2 plus an adjustment term that corrects for the fact that the two conditional distributions are obtained from different parts underlying the unconditional income distribution.

5.3.2 Application to the South African case

In order to obtain this derived Gini coefficient, Diaz-Bazan (2015) highlights the fact that consistent estimates of the distributions above and below threshold b are required.

Table 6 Gini coefficients at different thresholds. *Source*: Authors' calculations using NIDS and PIT (weighted)

Year	2011	2011 at 2014	2014
Combining data sets at different filing thresholds			
Threshold	R120,000	R250,000	R250,000
Overall	0.832	0.826	0.790
Below filing threshold (NIDS)	0.762	0.783	0.735
Above filing threshold (PIT)	0.367	0.326	0.349
Combining data sets at the 99th percentile			
Overall	0.826	–	0.791
NIDS	0.782	–	0.742
PIT	0.328	–	0.343

The availability of a high-quality household survey as well as a robust data set on tax records facilitates the application of this novel method to household survey data and tax administration data on South Africa. Table 6 reports the levels of inequality observed when the Diaz method is used to combine household survey data with tax administration data at the mandatory tax filing threshold for 2011 and 2014, respectively. The Gini coefficients are calculated using individual taxable income of adults 18 years and older. The decrease observed in overall inequality is much larger between 2011 and 2014 than what we observed when using household survey data alone in Sect. 5.1. From a total level of 0.83 in 2011, the Gini coefficient for taxable income decreased to 0.79. This extreme drop in inequality observed is driven by the fact that both conditional distributions report lower levels of inequality in 2014. Table 5 reported a slight decrease in inequality between the 2 years when using NIDS alone. However, in the household survey data, the decrease in the (conditional) Gini coefficient below the filing threshold was offset by a large increase in estimated inequality above the filing threshold. When using tax administration data above this threshold in order to explicitly account for the distortions caused by reporting bias and non-responses in the household survey data, we arrive at a much lower Gini for 2014. The decrease in overall inequality can also be observed when the 2011 data are assessed at the 2014 filing threshold albeit to a lesser degree. Applying the 2014 threshold to the 2011 data set increases dependence on household survey data at the top of the income distribution; however, it ensures comparability across the two years despite the shift in filing thresholds. In this case, inequality decreased from 0.826 to 0.79 when the two types of data sets are combined at the same filing threshold.

Additionally, Table 6 reports the Gini coefficients derived when household survey data and tax administration data are combined at the 99th percentile. In 2011, this reduces the Gini coefficient slightly from 0.832 according to the Diaz method to 0.826 when combined at the 99th percentile. In 2014, however, the difference is barely noticeable as the Gini increases from 0.790 according to the Diaz method to 0.791 when combined at the (arbitrarily) chosen 99th percentile. This increase is most likely driven by the large Gini coefficient above the filing threshold mentioned earlier that drives up the Gini coefficient measured in the household survey data. The

fact that the change in the Gini coefficient resulting from the change in thresholds is less significant in 2014 is most likely owed to the filing threshold of R250,000[34] being much closer to the 99th percentile in 2014 than it was in 2011. In the 2014 distribution, the filing threshold was at the 97.5th percentile. In 2011, the filing threshold was relatively lower at the 93.6th percentile. The 2014 filing threshold was at the 97.7th percentile of the 2011 distribution. When the 2011 data set is assessed at this threshold, the level of overall inequality observed is virtually the same as at the 99th percentile. Nonetheless, choosing an optimal threshold to combine the two data sets to explicitly account for reporting bias and non-response in the household survey data is necessary for consistent estimation of inequality. Therefore, the filing threshold is the preferred level at which the two types of data sets should be combined.

6 Conclusion

This paper takes advantage of the availability of high-quality household survey data and access to a unique set of tax administration data and uses a novel method in order to optimally assess levels of inequality prevalent across the entire population in South Africa. Our analysis of taxable income in the household survey data highlights a large fraction of zero income earners. This income group may earn income from other sources such as remittances or government grants but reports no taxable income.

The examination of the income distribution in the two underlying data sets has shown that there was a significant upward shift in taxable income in both years and across data sets. Using the Gini coefficient as a measure of inequality, we have shown that inequality decreased when using household survey data to assess the overall population. The NIDS data report a Gini coefficient of 0.823 for taxable income in 2011 and a Gini coefficient of 0.813 in 2014. Similarly, inequality measured above the filing threshold in the tax administration data has decreased over this period of time. Naturally, the measured inequality above the filing threshold is much lower than overall inequality and decreased from 0.367 in 2011 to 0.349 in 2014 according to the PIT data. Tax administration data below the filing threshold are unreliable and therefore cannot be used to assess the income distribution as a whole.

In our discussion of inequality across the different types of data sets, we have shown that there is an optimal threshold at which household survey data and tax administration data can be combined. This threshold minimizes dependence on household survey data while ensuring reliable information from tax administration data. The paper argues that this threshold should not be chosen arbitrarily and should rather be defined as the mandatory filing threshold. Detailed analysis of the different data sets has shown that household survey data in South Africa are unreliable above the filing threshold, especially in 2014. Therefore, household survey data from the NIDS survey are used to estimate the conditional distribution below the tax filing threshold and administration data on personal income tax are used for the estimation above the filing threshold. The results show that the decrease in inequality between 2011 and 2014 is even more significant in the estimation of the Gini coefficient of the thus combined income data.

[34] In PPP$, R250,000 is equivalent to $41,767 in 2011 prices (OECD 2017).

Applying the methodology introduced by Diaz-Bazan (2015), we estimate that the Gini coefficient of taxable income decreases from 0.83 in 2011 to 0.79 in 2014. A decrease in overall inequality is also observed when we combine NIDS data and PIT data at the 2014 filing threshold. Assessing the 2011 data sets at the 2014 filing threshold increases dependence on household survey data at the top of the income distribution for that year; however, it ensures comparability across the two years despite the shift in filing thresholds. Overall, choosing the tax filing threshold as the point of combining the two data sources is the best way to account for non-random reporting bias in the household survey data.

Despite the significant decrease in inequality between 2011 and 2014, our results are quite alarming for several reasons. Firstly, the extremely large fraction of zero earners indicate a strong dependency on non-taxable income sources. This holds pitfalls not only for the state and its capacity to generate revenue but also for the individuals at the bottom of the income distribution. Any form of non-taxable income they may receive will be small and less reliable than more structured forms of income that can be taxed.

There is a strong need for the state to address the severe deficits of individuals at the bottom of the income distribution. Seeing the macro-economic state of South Africa, however, it becomes clear that a continued heavy reliance on government grants may no longer be feasible. Instead, more structured ways of generating income need to be facilitated. Such solutions will require the government to introduce policies that induce growth and develop labour market policies to promote (self-)employment as well as skill development.

The novelty of the method by Diaz-Bazan (2015) used in this paper is only exceeded by the fact that we pioneer a type of analysis that combines household survey data with tax administration data in the South African context. However, more research can be done to identify the causes of the decrease in inequality observed in this paper and to determine their long-term feasibility in order to ensure continued efforts to transform South Africa into a more equal society.

Acknowledgements This paper was prepared for UNU-WIDER's Inequality in the Giants Project. The authors would like to thank participants at the UNU-WIDER Conference on Public Economics, held in July 2017 in Maputo, Mozambique, for their helpful comments and insights.

Appendix: Additional tables and figures

See Tables 7, 8, 9 and 10 and Fig. 6.

Table 7 NIDS attrition rate by deciles. *Source*: Authors' calculations using NIDS

Deciles by wave 1 HH income (2008)	Per cent refused in wave 4 (2014/2015) (%)
1	5.45
2	5.92
3	5.63
4	7.50
5	5.33
6	7.33
7	9.44
8	9.20
9	15.30
10	28.90
Total	100

Table 8 Statutory tax rates in South Africa. *Source*: SARS (2015)

2011	2014	Marginal PIT rates (%)
0–140,000	0–165,600	18
140,001–221,000	165,601–258,750	25
221,001–305,000	258,751–358,110	30
305,001–431,000	358,111–500,940	35
431,001–552,000	500,941–638,600	38
552,001 and over	638,601 and over	40

Table 9 Comparison of the information in the PIT and NIDS data sets. *Source*: Orthofer (2016) and NIDS

PIT data			NIDS equivalent	
	Item	Explanation	Variables	Notes
	Local interest	Local interest earned†	indi (dividends and interest) rnt	indi does not differentiate local and foreign dividends, cut-off of R22,300
+	Local capital gains	Excludes the basic exemption for capital gains (exclusion rate in 2010–2011: 75%)		
−	Local capital losses	Local dividends † ; rental profits/ losses; income from building societies; income from fixed period shares and deposits; royalties; foreign investment income (interest, dividends, capital gains/losses); gambling gains/losses	*Not covered*	
+	Other gains*		*Not covered*	
−	Other losses*		*Not covered*	
=	*Investment income incl. capital gains*			
+	Business profits	Profits/losses from unincorporated businesses or trades	emsprof_a	Profit after tax
−	Business losses		emsloss_a	
=	*Business income*			
+	Normal income	Local and foreign labour and pension income	em1inc em2inc swag prof ppen spen	does not differentiate between local and foreign benefits
+	Fringe benefits	Local and foreign lump-sum income, including remuneration and pension/ provident fund lump-sums	*Not covered*	
+	Lump-sum income		extr cheq bonu, em1prflm_a	does not differentiate between local and foreign benefits

Table 9 continued

PIT data			NIDS equivalent	
	Item	Explanation	Variables	Notes
=	*Labour income*‡			
=	*Gross income*	All incomes received by the individual		
−	Deductions	E.g. Taxes paid under pay as you earn; pension, provident or medical fund contributions. The exempted portion of interest income, all local dividends	em1dedmed_a em1dedpen_a em1deduif_a	
−	Exemptions		*Not covered*	
=	*Taxable income*	Taxable income used to determine the normal tax due (before any rebates and tax credits)		

* Asterisks refer to the subset of items under the respective SARS Code that are not mentioned separately in the table.
† Local interest below the threshold of R22,300 and local dividend income in its entirety are exempt from the PIT, and the accuracy of exempt incomes is not verified in the tax inspection process ‡ Employment income derived from taxable normal and lump-sum income (only taxable portion of normal and lump-sum income provided by SARS)

Table 10 Income brackets in the PIT data in PPP$. *Source*: Authors' calculations based on PIT Data (2011 and 2014)

Income group	2011		2014	
	Mean taxable income	Percentage of taxpayers (%)	Mean taxable income	Percentage of taxpayers (%)
0	0	4.53	0	2.34
Above 0–4189	2017	4.97	2112	3.77
4189–6284	5264	2.58	5255	2.03
6284–8379	7362	2.94	7360	2.27
8379–10,473	9470	3.34	9483	2.81
10,473–12,568	11, 609	4.59	11, 526	3.81
12,568–14,663	13, 638	4.82	13, 625	4.02
14,663–16,757	15, 694	5.30	15, 727	4.36
16,757–18,852	17, 805	4.63	17, 799	4.39
18,852–20,947	19, 893	4.04	19, 896	4.24
20,947–23,041	21, 971	3.84	21, 983	4.00
23,041–25,136	24, 116	3.67	24, 086	3.93
25,136–27,231	26, 176	3.66	26, 169	3.59
27,231–29,326	28, 264	3.32	28, 301	3.61
29,326–31,420	30, 375	3.09	30, 336	3.68
31,420–41,894	36, 309	13.73	36, 490	14.63
41,894–62,840	50, 522	13.12	50, 599	16.22
62,840–83,787	72, 194	5.63	72, 152	6.70
83,787–104,734	93, 156	2.91	93, 208	3.47
104,734–157,101	125, 683	3.11	125, 408	3.64
157,101–209,468	179, 284	1.01	178, 847	1.16
209,468–418,936	277, 072	0.90	277, 634	1.02
418,936–1,047,340	599, 155	0.24	601, 860	0.27
Above 107,340	1, 613, 911	0.03	1, 586, 025	0.03
N_weighted	5, 783, 360	100	5, 085, 060	100

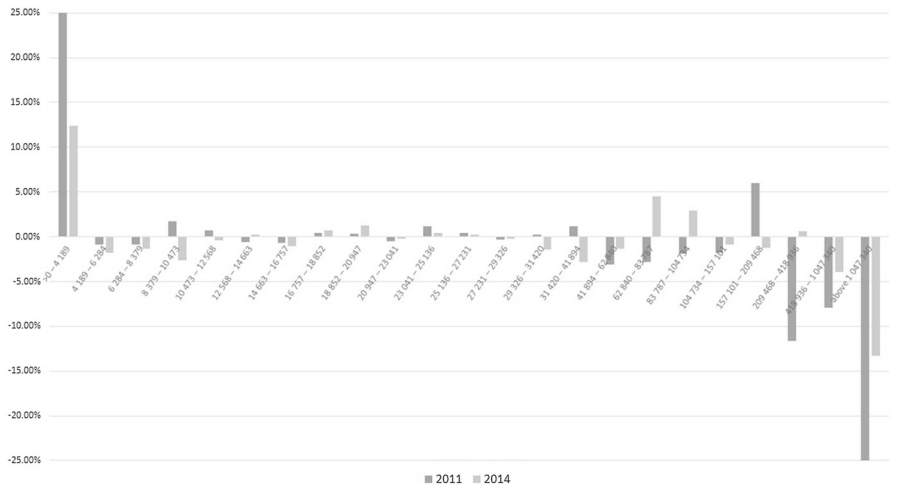

Fig. 6 Difference in estimated mean taxable income between NIDS and PIT by income bracket in PPP$

References

Alvaredo, F. (2011). A note on the relationship between top income shares and the Gini coefficient. *Economics Letters*, *110*(3), 274–277.

Alvaredo, F., & Atkinson, A. (2010). Colonial rule, apartheid and natural resources: Top incomes in South Africa, 1903–2007. CEPR Discussion Paper No. 8155.

Alvaredo, F., & Londoño, J. (2013). High incomes and personal taxation in a developing economy: Colombia 1993–2010. Tech. rep., CEQ Working Paper.

Atkinson, A. B. (2007). *Measuring top incomes: Methodological issues* (pp. 18–42). Oxford: Oxford University Press. (ch. 2).

Atkinson, A. B., Piketty, T., & Saez, E. (2011). Top incomes in the long run of history. *Journal of Economic Literature*, *49*(1), 3–71.

Burkhauser, R. V., Feng, S., Jenkins, S. P., & Larrimore, J. (2012). Recent trends in top income shares in the United States: Reconciling estimates from March CPS and IRS tax return data. *Review of Economics and Statistics*, *94*(2), 371–388.

Dell, F. (2005). Top incomes in Germany and Switzerland over the twentieth century. *Journal of the European Economic Association*, *3*(2–3), 412–421.

Diaz-Bazan, T. V. (2015). Measuring inequality from top to bottom. World Bank Policy Research Working Paper, 7237.

Finn, A., Leibbrandt, M., & Levinsohn, J. (2012). Income mobility in South Africa: Evidence from the first two waves of the National Income Dynamics Study. NIDS Discussion Paper, 05.

Garbinti, B., Goupille-Lebret, J., & Piketty, T. (2016). Income inequality in France, 1900–2014: Evidence from Distributional National Accounts (DINA). Tech. rep., WID. world Workong Paper Series.

Hundenborn, J., Woolard, I., & Leibbrandt, M. (2016). Drivers of inequality in South Africa. Tech. Rep. 194, Southern Africa Labour and Development Research Unit.

Inchauste, G., Lustig, N., Maboshe, M., Purfield, C., & Woolard, I. (2015). The distributional impact of fiscal policy in South Africa. SALDRU Working Paper.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in UK income inequality. *Economica*, *84*(334), 261–289.

Jorda, V., & Niño-Zarazúa, M. (2016). Global inequality: How large is the effect of top incomes? Tech. rep., WIDER Working Paper.

Lakner, C., & Milanovic, B. (2016). Global income distribution: From the fall of the Berlin Wall to the Great Recession. *World Bank Economic Review*, *30*(2), 203–232. https://doi.org/10.1093/wber/lhv039.

Leibbrandt, M., & Finn, A. (2012). Inequality in South Africa and Brazil—Can we trust the numbers? CDE Insight, Centre for Development an Enterprise. https://www.africaportal.org/publications/inequality-in-south-africa-and-brazil-can-we-trust-the-numbers/

Leibbrandt, M., Finn, A., & Woolard, I. (2012). Describing and decomposing post-apartheid income inequality in South Africa. *Development Southern Africa*, *29*(1), 19–34. https://doi.org/10.1080/0376835X.2012.645639.

Levy, B., Hirsch, A., & Woolard, I. (2015). Governance and inequality: Benchmarking and interpreting South Africa's evolving political settlement. ESID Working Paper No. 51.

Lustig, N. (forthcoming) The missing rich in household surveys: Causes and correction approaches. CEQ Working Paper.

Morelli, S., Smeeding, T. M., & Thompson, J. P. (2014). Post-1970 trends in within-country inequality and poverty: Rich and middle income countries. Center for Studies in Economics and Finance Working Papers, 356.

National Planning Commission. (2011). National Development Plan 2030. Our future—Make it work. https://www.gov.za/sites/default/files/gcis_document/201409/ndp-2030-our-future-make-it-workr.pdf

National Treasury and South African Revenue Service (SARS). (2015). Tax Statistics. *2015*. https://www.sars.gov.za/AllDocs/Documents/Tax%20Stats/Tax%20stats%202015/Tax%20Statistics%202015.pdf

OECD. https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm. (Website) PPPs and exchange rates. Accessed Nov 23, 2017.

SALDRU. (2015a). National Income Dynamics Study (NIDS) 2010–2011, Wave 2 [dataset]. Southern Africa Labour and Development Research Unit, Version 3.1.

SALDRU. (2015b). National Income Dynamics Study (NIDS) 2014–2015, Wave 4 [dataset]. Southern Africa Labour and Development Research Unit, Version 1.1.

SARS. (2016a). Personal Income Tax (PIT) Data 2010–2011 [dataset]. South African Revenue Service.

SARS. (2016b). Personal Income Tax (PIT) Data 2013–2014 [dataset]. South African Revenue Service.

Schiel, R., Leibbrandt, M., & Lam, D. (2016). *Assessing the impact of social grants on inequality: A South African case study* (pp. 112–135). Contemporary Issues in Development Economics New York: Palgrave MacMillan. (ch. 8).

Statistics South Africa. (1993–2017). Quarterly publication on Gross Domestic Product (GDP). Statistics South Africa P0441.

Van Der Weide, R., Lakner, C., & Ianchovichina, E. (2016). Is inequality underestimated in Egypt? Evidence from house prices. World Bank Policy Research Working Paper.

Wittenberg, M. (2016). Wages and wage inequality in South Africa 1994–2011: Part 1-Wage measurement and trends. *South African Journal of Economics*, *85*(2), 279–297.