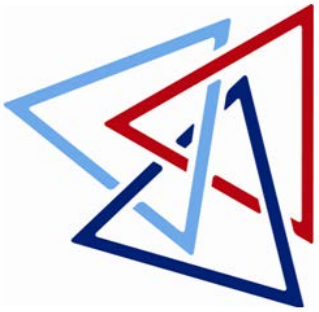
REDI3x3 Working paper 41 October 2017



Measurement of earnings: Comparing South African tax and survey data

Martin Wittenberg

Abstract

Comparing earnings in the tax assessment data to those in the QLFS, it appears that earnings of employees in the QLFS are underreported by perhaps 40%, with bigger gaps near the top of the distribution. Benefits and annual bonuses contribute substantially to the gap. In the case of self-employment incomes it is also the case that high earnings are missing or underreported in the QLFS, but the tax data seems to miss many mid- and low-income self-employed earners. These differences make sense when one considers the incentives for reporting accurately to SARS versus to Statistics South Africa. These errors mean that earnings inequality as measured by the Gini coefficient is probably underestimated in the surveys by three percentage points.

The **Research Project on Employment, Income Distribution and Inclusive Growth** is based at SALDRU at the University of Cape Town and supported by the National Treasury. Views expressed in REDI3x3 Working Papers are those of the authors and are not to be attributed to any of these institutions.

Measurement of earnings: Comparing South African tax and survey data*

Martin Wittenberg
School of Economics, SALDRU and DataFirst
University of Cape Town
South Africa

1 Introduction

How reliable are South Africa's household surveys when it comes to recording labour market earnings? This is an important question for many policy debates. For instance the proposed introduction of a national minimum wage led to heated discussions about the likely impact of setting it at different levels (Finn 2015, Seekings and Nattrass 2015a, Isaacs 2016). Determining how many workers (and firms) would be affected is difficult if the earnings reported in the Quarterly Labour Force Surveys (QLFSs) are not accurate. Many analysts assume that earnings are underreported (e.g. Seekings and Nattrass 2015b, p.57). Indeed there is a systematic mismatch between firm data and individually reported earnings, with average wages in the firm surveys considerably higher than those reported in the QLFS (Wittenberg 2014, p. 38).

Inaccuracies in the reported earnings will also affect poverty and inequality estimates. One of the key questions is whether any under-reporting is likely to be more pronounced at the bottom or the top of the earnings distribution. In order to get any traction on these questions it is necessary to get a source of data that is not subject to the reporting biases of the household surveys. In this paper I will utilise South African tax data, in particular the tax assessments. That brings with it its own set of challenges. In particular we need to take account of the fact that most tax payers are not required to file returns. This will lead to a selected sample at lower earnings levels. Nevertheless we should have a much better coverage of high income earners than is the case for any of the surveys. Comparing our results to those from the QLFSs will throw some light on what sort of underreporting we are likely to see at the top of the distribution.

The plan of the discussion is as follows. In the next section I discuss some of the known measurement issues in the South African surveys. In section 3 I discuss in more detail the nature of the data and how I choose to define the sample and variables. The Methods section outlines how I intend to tackle the problem of comparing two very different distributions. Section 5 presents the results and these are subsequently discussed. I draw some lessons for the quality of the data in the conclusion.

*I would like to thank REDI3x3 for funding this research and for giving me access to the data. Many thanks to Elizabeth Gavin for talking me through some of the intricacies of the data. I would also like to thank an anonymous referee for helpful comments. Ingrid Woolard, Kezia Lilenstein and Andrew Kerr provided useful feedback on some parts of this work. None of them are responsible for any errors in the analysis.

2 Literature Review

Wittenberg (2017a, 2017b) provides an overview of many data quality issues associated with the earnings information in South Africa's household surveys. A key problem is missing information. Richer individuals are more likely to refuse to answer the earnings question. Some of them give their earnings bracket to a follow-up question but significant numbers reveal nothing. This problem is compounded by the fact that outright refusal to participate in the survey (i.e. unit nonresponse) is also higher in affluent areas. In order to deal with these issues the survey organisations and analysts employ weighting adjustments and/or imputations (Wittenberg 2017a), but these approaches rely on the fact that the observed information is a reasonable guide to the type of data that has been missed due to nonresponse.

Missing data is not the only problem. Measurement error is likely to be important too. In most surveys individuals are asked about gross income (Wittenberg and Pirouz 2013). For instance, in the QLFS the question was asked in two parts (Statistics South Africa 2010, Q 5.2 & Q 5.4a):

> "In your main job, what is the easiest way for you to tell us your wages or salary before taxes or any other deduction? Would it be ... monthly, weekly, fortnightly, daily, hourly, annually?" and then "What is your (choose one) annual/ monthly / weekly / daily / hourly wage or salary before deductions? (Include tips and commissions)"

There is lots of scope for misreporting here. The figure which is likely to come most immediately to mind is take-home pay. To add back the tax plus medical aid and pension contributions may be too cognitively demanding for many respondents. According to Seekings and Nattrass (2015b, p.57):

> "Evidence from NIDS suggests that most workers do understand the difference between gross and take-home pay, but there is no direct evidence as to whether they accurately gauge this difference when reporting their gross earnings. Van der Berg found that government payroll indicate public sector teacher's gross earnings were 40 percent higher than household surveys suggest"

Wittenberg (2014, p.44) tried to reconcile the earnings recorded in the household surveys with those obtained from firm surveys. He showed that aggregate earnings in the mining sector were, coincidentally, also 40% higher when using the firm data (which arguably should be quite accurate for mining) than when using the Quarterly Labour Force Survey. The problem of correctly adding back tax and other deductions to derive an accurate "gross earnings" figure may be compounded by the fact that in a number of cases the survey information is supplied by a proxy respondent and not the person earning the income themselves (Daniels 2012).

In the case of self-employed individuals the measurement problems are likely to be even worse. The QLFS question was again in two parts (Statistics South Africa 2010, Q 5.6 & Q 5.7a):

> "What is the easiest way for you to tell us your earnings after expenses? Would it be ... monthly, weekly, fortnightly, daily, hourly, annually?" follwed by "What are your (choose one) annual/monthly / weekly / daily earnings after expenses?"

In this case individuals are required to have a good sense of the difference between business and personal expenses. Performing the required mental arithmetic is likely to be difficult, particularly for informal sector operators.

How big are the measurement errors likely to be? Wittenberg (2014) tried to adjust for differences in coverage between the household surveys and firm surveys, non-response by high

earners, hefty underreporting of earnings by those in the sample, possible underreporting of employment by firms and at the end still had an unexplained gap of around 15%. He concluded:

> "It is hard to know what to make of it. What is really required is the sort of hard data (e.g. PAYE information) that would give us both accurate aggregate earnings information as well as good measures of its distribution." (p.47)

At that time tax information was not available for research of the sort envisaged there. Since then the value of administrative data for answering pressing policy questions has been increasingly recognised. For instance linked firm and individual tax data has been made available by the South African Revenue Service via the National Treasury to investigate a range of questions (Pieterse, Kreuser and Gavin 2016) from churn in the labour market (Kerr 2016) to the impact of importing on firm performance (Edwards, Sanfilippo and Sundaram 2016). A sample of personal income tax assessments was also made available via National Treasury to REDI3x3. Those data have been used to investigate inequalities in income and wealth (Orthofer 2016). It is the dataset that we will be using in the empirical work below.

The advantage of tax data over the survey information is that there are serious consequences to "non-response" in relation to tax filing, at least above the relevant tax thresholds. Furthermore for many employees the earnings are directly reported by their employers, through the "Pay as you earn" (PAYE) system. This reports also on all the benefits such as medical aid and pension contributions for which the employee becomes liable for tax. Consequently the tax data is likely to provide a better measure of gross earnings than individuals' self-reports. Nevertheless there are also limitations to the tax data. Individuals have an obvious interest in minimising the taxable amounts that they declare. Consequently we would expect those types of earnings that are more difficult for the authorities to monitor (e.g. certain forms of self-employment incomes) to be less well reported. Furthermore we know that many rich people will actively structure their remuneration portfolio in ways that will minimise their obligations. This is before one considers outright avoidance.

One additional key limitation in the South African context, discussed by Orthofer (2016), is the high filing threshold of R120 000 per year. This means that the assessments data is missing the bulk of all earners. Orthofer (2016) got around this difficulty by imputing a truncated log-normal distribution for the bottom tail of the tax data and splicing in a Pareto distributed upper tail to the survey distribution. Given these adjustments she found the income distribution from the survey and tax data to be remarkably similar:

> "the resulting measures of income inequality coincide almost perfectly between the two sources: one percent of the population earns 16-17 percent of all incomes; ten percent earn 56-58 percent" (p.3)

In this paper I want to revisit this finding and concentrate in particular on the earnings from work. Given the discussion about the different sources of error, I will want to contrast earnings from "regular work" with earnings from own account activities.

In the empirical work I will use Pen's Parade of many dwarves and a few giants (Pen 1971, pp.48ff) as a key organising device for thinking about the distribution of income. This has not often been done in South Africa (exceptions are Tregenna and Tsela 2012, Wittenberg 2012) so it is useful to outline the core idea. Pen's parade proceeds by lining up individuals from poorest to richest and then lets them walk past within an hour. Unlike an ordinary parade, however, individuals are scaled relative to their height, so that the individual earning the average income, is of average height. Wittenberg (2012) discusses what this looks like in South Africa given the per capita household income data from the 1993 Project for Statistics on Living Standards and Development:

"After a minute, we have already seen 650 000 people streaming by, but we might not have noticed anything: that is because these people are all really tiny, hardly visible at all – after the first minute the height of the dwarf scooting by is only 1.4 cm. After 2 minutes and a total of 1.3 million people later, the height is still only 4cm; after a full 10 minutes we see dwarves of 20 cm. After 30 minutes and half way through our parade we still don't look the marchers in the face; they don't even reach up to my belt, as they are 60 cm in height. At the three quarters of an hour mark we have individuals who are of recognisably adult stature, although they still don't measure up to me yet. Their height is 156cm. Two minutes later, however, an individual of my height has shot past and we are suddenly among giants. At the fifty minute mark we have a 2.6m individual and five minutes later it is a 5m giant. One minute before the end of the parade we have 650 000 people still to come, but they are now around the height of a four story building at 11m. In the last few seconds we have individuals taller than Table Mountain and some perhaps even poking out of the atmosphere. I can't tell because the super-super-rich are not properly represented in our surveys."

We will revisit this exercise, but for earnings, below.

3 Data

I use two types of data. The tax assessments data for 2011 and the four waves of the Quarterly Labour Force Surveys corresponding to that period, i.e. 2010 quarters two, three and four and 2011Q1[1].

3.1 Tax assessments 2011

The assessments data comes in two forms: there is a 20% sample of all Personal Income Tax (PIT) assessments excluding those with a taxable income of above R10 million (472 individuals) and there is aggregate information on the top 472 income earners. The key variables that provide information on how remuneration is structured are shown in Table 1.

The dataset also includes total taxable income and the tax that is liable. Besides the income information the dataset also includes the age of the taxpayer, their gender, marital status and number of dependents. This, of course, is another limitation of tax data – there are very few usable covariates.

3.2 Quarterly Labour Force Survey

Labour market information has been collected through nationally representative surveys at least since the 1993 Project for Statistics on Living Standards and Development. The instruments through which earnings are recorded have gone through several changes since those early days (Daniels 2012, Wittenberg and Pirouz 2013). Since 2008 the Quarterly Labour Force Survey is the main instrument through which Statistics South African collects information on labour market outcomes. The questions through which earnings are elicited have already been reported. What is important to note, however, is that individuals can only report **either** earnings from employment **or** the returns from own-account activities, but not both. In practice (as the tax data shows) individuals who work for an employer may still earn some money on the side,

[1] The tax year goes from March to February, so there is one month of 2011Q1 which falls outside the tax year. Andrew Kerr pointed out to me that most of the fieldwork is done in February, so this mismatch is not all that important.

Variable	Definition
Income variables (code 36)	
income3601	income (PAYE). In our dataset it also includes overtime. Less helpfully it also includes pension and retirement annuity payouts.
income3605	annual payment
income3606	commission earned
income3615	director's income
income_36range	this type of income includes arbitration awards, independent contractors, labour brokers, restraint of trade, foreign incomes both taxable and non-taxable, foreign pensions etc.
Allowances (code 37)	
travel3701	travel allowance
shares3707	share option exercised
allotherallowances	including entertainment allowance, public office allowance, uniform allowance, tool allowance, computer allowance, telephone allowance, subsistence allowance, employees broad based share scheme, foreign allowances. It is not clear whether this code also includes fringe benefits other than medical aid.
Fringe benefits (code 38)	
medical3810	medical aid contribution paid on the taxpayer's behalf
Lump sums (code 39)	
income_39range	includes gratuities, pension fund lump sum payouts upon resignation or retirement, special remuneration, insurance gains, unclaimed benefits
Profit/Loss (codes 1 to 34) **and Investment income** (code 42)	
profit_1_34range	aggregate profits reported for all types of economic activitiy, e.g. farming, mining, food production, textiles
loss_1_34range	aggregate of losses reported for all types of economic activitiy, e.g. farming, mining, food production, textiles
income4201	local interest earned
profit_42range	includes dividends, share income, rental income, royalties, foreign interest earned, capital gains, sporting income, gambling
loss_42range	includes rental losses, capital losses, foreign investment losses
Deductions (code 40)	
currpens4001	current pension fund contribution
currracontr4006	current retirement annuity contribution
medexpen4008	total medical expenses
travelexpen4014	travel expenses – fixed costs
travelexpen4015	travel expenses – actual costs
allotherdeductions	includes expenses on tools, entertainment, home office, depreciation, allowable accountancy fees

Table 1: Information in the Tax Assessments Data

or individuals who may have substantial business interests may still receive payments through the PAYE system from firms. It is also worth noting that the earnings data from the QLFSs is released separately from the other variables under the label of the South African "Labour Market Dynamics" study. The information from all four quarters of any year are released together. The versions that we will use in this study are those released by DataFirst as part of PALMS - the Post-Apartheid Labour Market Series (Kerr, Lam and Wittenberg 2016). That version enables us to break up the information back to the constituent quarters and, in this case, align the periods of the data better. It is worth noting also that the earnings data arrives from Statistics South Africa with imputations for bracket responses and other types of missing data. At this stage I'm not able to discuss how that was done.

We have already seen some of the weaknesses of the survey data: the incentive for respondents to get the answer to the income question right is much lower than it is for tax data. Indeed there are no penalties to not responding at all. It is worth noting that even in the absence of nonresponse and measurement error, survey data would battle to accurately capture the upper tail of the income distribution – simply due to the fact that with the current sampling strategies one would expect to get at most a handful of the super-rich into the sample.

On the plus side, the fact that the survey information does not go back to the tax authorities may prompt some individuals to report their incomes more freely or honestly than they might be inclined to if they knew that they were going to be taxed on their answers. Another potential strength of the survey information is that one has many other variables as potential covariates. This means that much research will of necessity still need to use survey data as their baseline. There is therefore considerable interest in getting a fix on how accurate the earnings information is.

3.3 Making the data comparable

3.3.1 Sample definition

One of the problems of the tax assessments data is that with the data available to us we cannot separate out regular earnings from pension payments. In order to weed out most of the problematic cases I restricted the age range to individuals aged twenty to sixty. Additionally, the sample of individuals in the assessments data with incomes below the filing threshold of R120 000 will be selected in complicated ways. In particular they will include individuals who have incentives to claim bigger deductions than they would get through the normal PAYE system and "provisional taxpayers", i.e. individuals with additional incomes. One way of dealing with this mismatch would be to simply eliminate these individuals from the sample also. The main technique that we will use to compare the distributions (to be discussed below) does not require us to do this, but it is worth noting this issue upfront.

Wage earners The tax data does not provide the neat division into "wage earners" and "employers plus own account workers" (the "self-employed") that is a core organising principle of the labour market data. Given the fact that people report different types of earnings, how could one get a categorisation that broadly resembles this division? I choose to regard someone as a "wage earner" if regular income (income3601 in Table 1) makes up more than half of all taxable income. In cases where taxable income is lower than regular income (due to losses and other deductions) I regard someone as "wage earner" if the taxable income is still positive; once business losses are bigger than earned income, I would be inclined to think that the primary activity of the individual is business rather than working for an employer.

Category (Taxable Income Range)	Earners	Self-Employed	Rentiers	Prop Earners	Prop Self Emp
1000000-9999999	29415	25860	35	.532	.468
500000-999999	169270	34000	110	.832	.167
250000-499999	599965	62595	440	.905	.094
120000-249999	1513430	89980	1225	.943	.056
60000-119999	1076740	79605	1660	.930	.069
0-59999	542655	291900	6975	.645	.347
<0	0	50085	635	0	.987
Note: Data weighted to give estimated population counts					

Table 2: Breakdown of the Assessments Data

The self-employed Given the dichotomous definition in the Quarterly Labour Force Surveys, anyone who isn't a wage earner should be self-employed, provided they are "working". Of course it is possible for somebody to be earning an income and be neither a wage earner nor self-employed. Pensioners should be mainly excluded by our age criterion, but there will be some early retirees as well as people drawing disability benefits. There are also likely to be at least some pure rentiers. How would one identify the latter in these data? Anyone whose predominant source of income is interest (income4201 in Table 1) arguably falls into that category. Undoubtedly there will also be a number of rentiers receiving the bulk of their income in dividends and rents (profit_42range). Unfortunately we cannot distinguish individuals who actively manage their portfolios (e.g. the "day traders") from those who passively live off portfolios managed by others. Nor can we divine how such individuals would choose to categorise themselves if they were to be confronted with the QLFS instrument – would they consider themseslves to be working or not? In the absence of better alternatives, I exclude individuals where interest income (income4201) exceeds 90% of taxable income, but do not exclude anyone else.

The impact of these definitions In Table 2 we show how the tax assessments sample divides up according to these criteria. We have organised the table by different ranges of taxable income. It is evident that the "self-employed" are represented heavily at the top of the distribution and then again near the bottom, i.e. well below the compulsory filing threshold (for PAYE earners) of R120 000. It is also evident that there are not many pure rentiers, i.e. individuals that we have effectively categorised as not working. For most individuals wages (income3601) are the predominant form of income.

Some additional caveats There are some differences in coverage between the samples that we cannot address. There are some South African taxpayers that are not currently resident in the Republic and there are some employees of foreign agencies (e.g. the World Bank) currently in South Africa that will not need to pay South African tax. Theoretically these individuals will be covered in different ways in the datasets. There are unlikely to be many of these cases, but at least a few (e.g. South African executives currently working abroad) may show up as very high earners in the tax data.

3.3.2 Variable definition

The QLFS concept of gross earnings doesn't translate neatly in tax categories. For the sample of "wage earners" I will be working with the following:

- **regular earnings**: i.e. income3601 in Table 1
- **cash earnings**: I add annual payments (income3605), commissions (income3606), director's income (income3615), other types of cash income (income_36range) as well as share options that were exercised (shares3707) to regular earnings
- **total earnings**: to the cash earnings I then add medical aid benefits (medical3810), all other allowances (allotherallowances) plus travel allowances (travel3701). I do however deduct travel expenses from the latter (travelexpen4014+travelexpen4015).

One problem is that the QLFS question for wage earners is about wages from "the main job". There is no guarantee that either the "regular earnings" reported through the PAYE system or any of these additional streams of income all relate to the same "job". That is a potential limitation that I cannot overcome, however, other sources of income (such as profits) are not included.

In the case of earnings for the "self-employed" there is no analogous way to proceed. Instead I construct a "gross income" variable from taxable income as follows:

- **gross income:** taxable income plus medical expenses incurred (medexpen4008), current pension contributions (currpens4001) and retirement annuity contributions (currracontr4006)

4 Methods

4.1 Reverse Pen's Parade

The key tool through which I will compare the tax data and the survey evidence is a "reverse Pen's parade". Like the original parade it involves ordering the population on earnings, except that in this case we start with the richest and proceed to the poorest. In some implementations (e.g. Tregenna and Tsela 2012, p.41) Pen's parade is represented as a graph in which the y co-ordinate is given by income and the x-coordinate is relative position in the population (from $\frac{1}{N}$ for the poorest to 1 for the richest). I not only reverse the rank order, but also plot this graph against absolute rank (from 1 for the richest to N for the poorest). I do this because I run the parade for the tax data alongside the one from the QLFS, and we know that the lengths of the two parades are very different. My approach will produce valid comparisons as long as we have data points that "represent" the same population. Obviously we will need to use weights, where necessary (i.e. in the surveys) adjusted for non-response. The key question that I want to investigate is how different these levels are: how much richer is the person in position X in the tax data than the person in the same spot in the survey data? (Or *vice versa*)

The tax data will be ranked in terms of **regular income** (i.e. income3601). The other types of income are not monotonic functions of regular income, i.e. cash earnings can vary quite a lot for individuals with similar regular incomes. In order to present interpretable output, I will calculate the average cash income or average total earnings in a neighbourhood of a particular regular income value. This is done by means of a nonparametric (local linear) regression.

In the case of the survey evidence we need to decide whether to pool the data from all four quarters (2010Q2 to 2011Q1) or whether to run separate comparisons by quarter. My first analysis will therefore be designed to establish whether it is fair to pool the QLFS data. I show below that this is, indeed, reasonable. Since the weights are designed to make the data for each quarter representative of the entire population, I divide these by four when I pool the four quarters.

4.2 Dealing with the top 472 earners

We have an immediate problem at the front of the tax data parade, since we do not have individual level information on the richest few individuals. In particular, we don't even know whether any of them would qualify to be "wage earners" by the definition that I used earlier. We do know (from the aggregate information) that more than R800 million was earned collectively by them in income3601, so there are likely to be at least a couple that would meet our criteria. In the absence of additional information I assume that the number of "employees" in this group is proportional to the proportion of this regular income of all taxable income in this group. It gives a count of perhaps 51 individuals. Of course not all of the regular income would have accrued just to this group. But then some of the other types of earnings (e.g. allowances and benefits) would accrue to them also. It is hard to know which type of error will be worse. Finally I reduce this by a quarter to 38, since around a quarter of the richest are above sixty.

Similarly we estimate the number of pure rentiers proportional to local interest in total taxable income, giving an estimate of 13. The balance would then be "self-employed". Again we reduce this by a quarter to give us 306 "employers or own account workers" among the super-rich. For the parade of the "self-employed" I calculate gross income among the top using the same generic formula I outlined in section 3.3.2, i.e. taxable income plus medical expenses, but I stripped out regular income and interest income from the taxable income aggregate and scaled medical expenses, pension and retirement annuity payments proportional to the number of people in this category.

4.3 Aligning the tax and survey parades

Both the "tax parade" and the "survey parade" are based on sample data – this means (in the case of tax data) that we see only every fifth person going by. In the case of the survey data the weight (when we pool the data) is of the order of 170, i.e. we see on average every 170th earner. This means that the actual data points on which we want to compare the two distributions will be sparse and will not be perfectly aligned. In order to get measurements for both distributions at precisely the same rank values, I run a nonparametric regression (local linear regression) of income on rank to get suitably interpolated values at the same grid point values. I use the Stata "lpoly" command with the default plug-in bandwidth selection for this purpose.

5 Results

5.1 Wage-earners

An initial check on whether or not it is possible to pool the survey data is given by Figure 1 which suggests that except perhaps right at the top of the distribution, it isn't particularly problematic to pool the four waves of the QLFS. Given the small number of observations at the top of the distribution (note the log scale) it isn't altogether surprising that there should be some differences. It is perhaps more surprising that the initial level is not more different for the three waves of the 2010 QLFS – whereas the 2011Q1 series starts at a noticeably lower level. A closer look at the data shows that the top annual income in the QLFS was R4 800 000 which was recorded eight times in the data. Three of these cases were of the same individual from the Eastern Cape and another two were due to one woman in Gauteng. Given the rotating nature of the panel some repeat data is to be expected. However there were therefore still five distinct individuals who recorded this precise income at some stage in 2010[2]; but this amount was not

[2] Indeed there were an additional two distinct individuals with that income in 2010Q1.

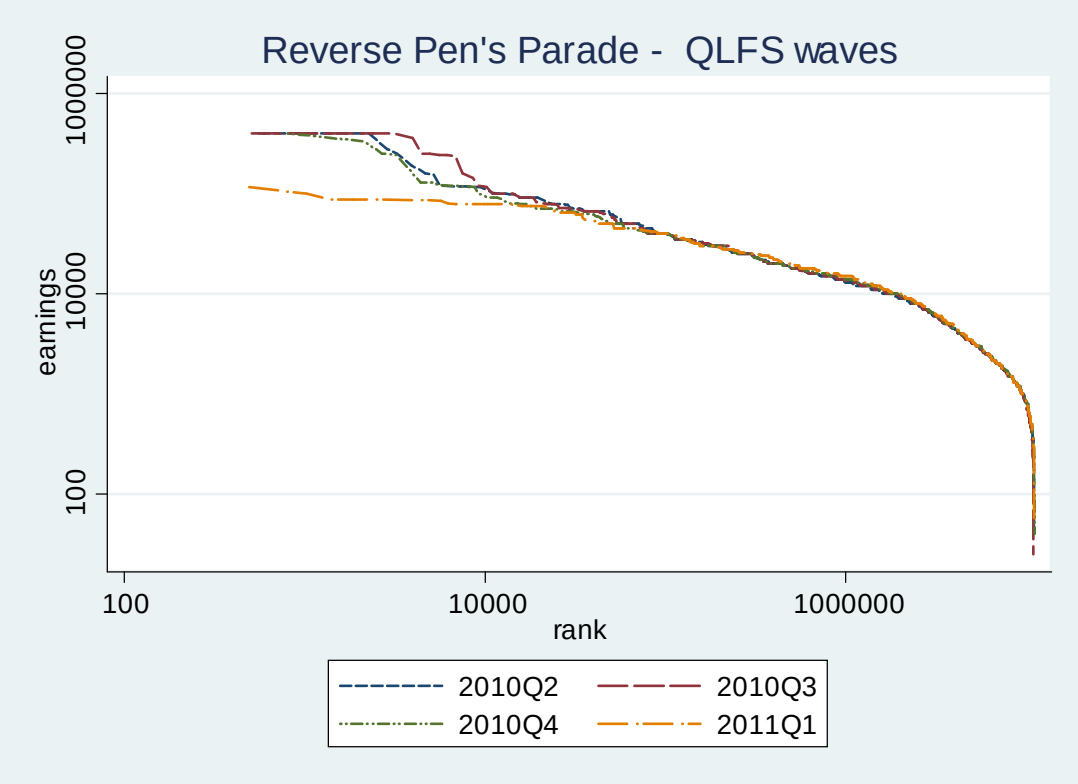


Figure 1: The Pen's parades from the four waves of the 2010/11 tax year

recorded in the first quarter in 2011 or indeed at any other time in that year: the top income in all of 2011 was R3 960 000. This suggests that the multiple occurrence of the same number may be due to the way missing data were imputed in 2010 (since the earnings data from those surveys were released together).

5.1.1 The survey data compared to regular earnings

The relationship between the "tax parade" and the "survey parade" is shown in Figure 2. There are several interesting features:

- The earnings right at the top of the tax distribution are higher than in the surveys. Even excluding any "employees" among the top 472 tax payers, there are 9 individuals in the tax assessments data (representing 45 individuals in the population) with annual earnings higher than R4.8 million. Given our imputations for the very top, we think there are perhaps 83 individuals with earnings above this level.
- The flat portion at the start of the "tax parade" represents income imputed for the 38 top earners deemed to be among the top tax payers. The much more extensive flat portion at the start of the "survey parade" is due to the individual-quarter records with exactly R4.8 million in earnings. The weight of these 8 records is 1553.

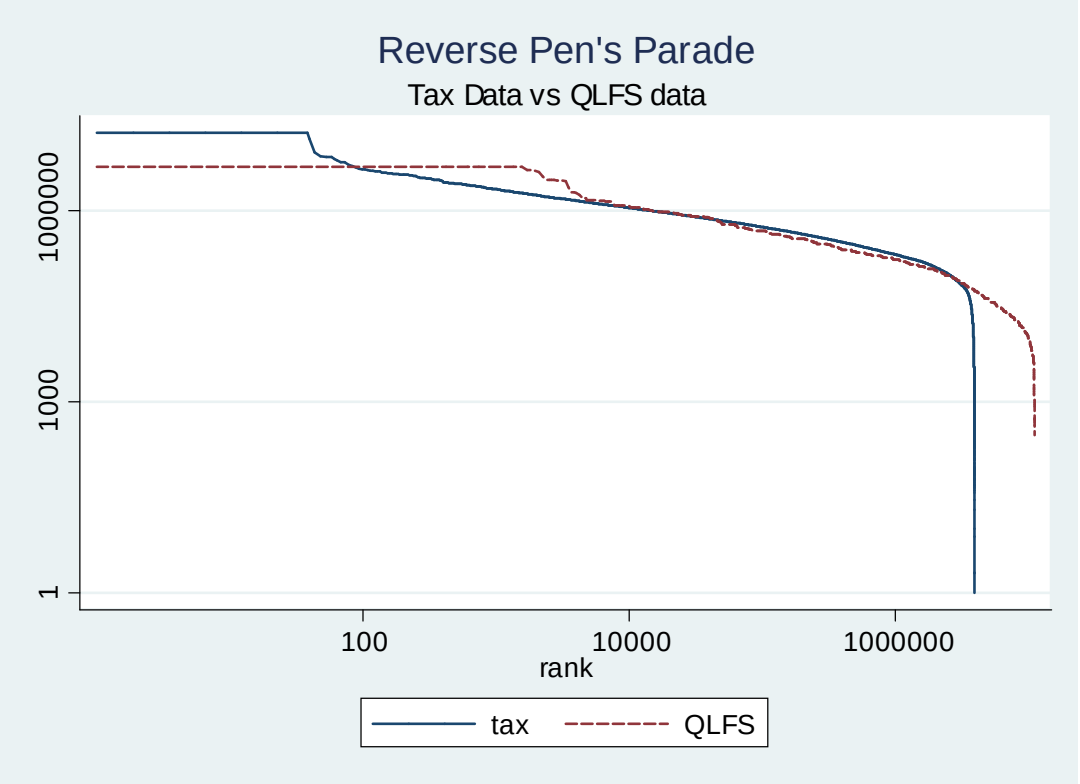


Figure 2: The relationship between "regular income" (income3601) in the tax data and earnings of employees in the survey data

- There are a couple of other high earners in the survey data which means that once the two lines cross (around rank 83) the individuals in the survey parade are actually taller than the ones in the tax parade right up to rank 44520 (represented by the 213-th individual in the sample data). This corresponds to an income of R713 000 per annum, i.e. just under R60 000 per month.
- For the rest of the parade, the "regular income" in the tax data is somewhat higher than the income in the survey data, until around rank 2 700 000 corresponding to earnings of around R90 000 (R7 500 per month). This, of course, represents earnings below the compulsory filing threshold, so the tax data becomes increasingly anomalous.

While the mismatches between the two distributions are interesting, it is astounding that they are not much bigger. By the looks of it the survey data misses out on a handful of really high income earners, but otherwise it tracks the distribution of "regular earnings" remarkably well. The take-home message of Figure 2 is that to a first approximation the survey earnings question for employees captures "regular earnings", i.e. income3601 in the tax data.

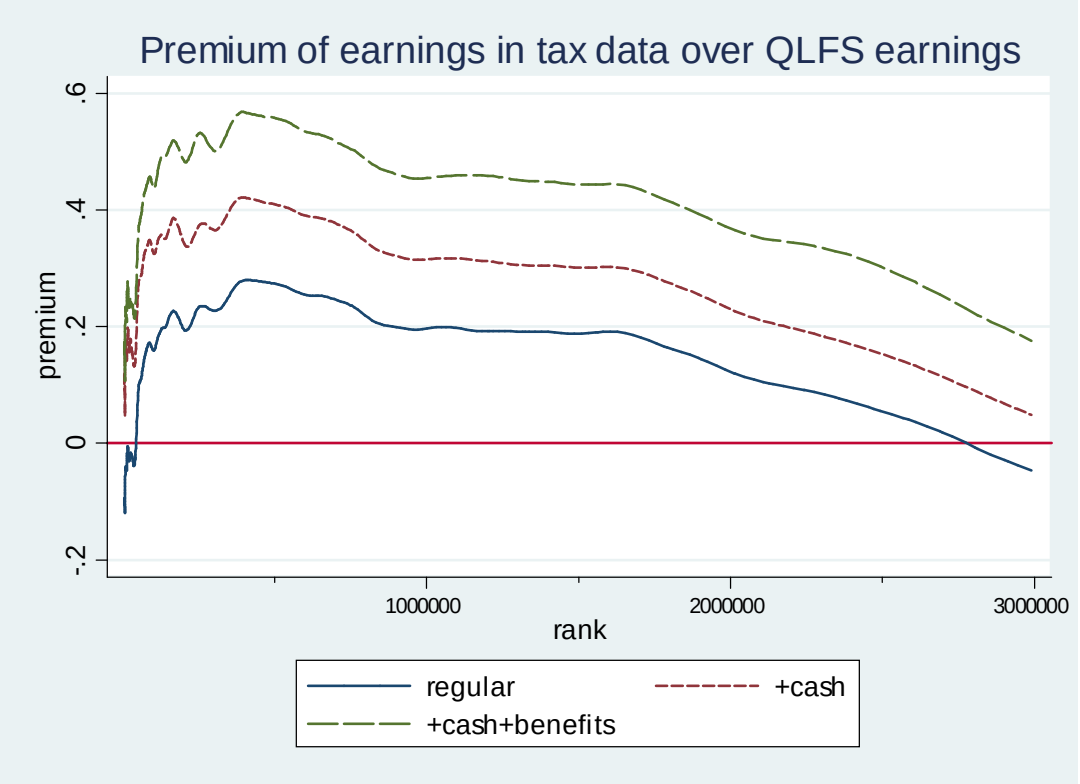


Figure 3: The relationship between survey earnings and earnings in the tax data for employees

5.1.2 The relationship between regular earnings and total earnings

In Figure 3 we show that the apparently small gap between "regular earnings" and the survey data amounts to a fairly constant 20% gap over the bulk of the distribution where the tax assessments data is accurate. The premium is defined as $\frac{(tax-survey)}{survey}$. For reference the income3601 of the individual with rank 2 000 000 is R 128 000, which is just over the filing threshold. In Figure 3 it is also clear that the gaps are much larger once one adds in additional cash payments (in particular annual lump sums) and benefits (particularly medical aid). It also seems evident that the gaps are larger at the top of the distribution (the front of the parade), with the exception of the first 40 000 where, as we showed in Figure 2, the survey income actually seems to exceed the one in the tax data.

Looking at the tax data by itself, we show in Figure 4 the premia of cash earnings and total earnings over regular earnings, e.g. we graph $\frac{(cash-regular)}{regular}$ against rank in the distribution. This figure suggests that for the bulk of the distribution, cash lump sums (e.g. annual bonuses) increase regular earnings by around 10% on average – but they make a much bigger difference for high earnings individuals. Once benefits are added, the average total package is between 20 to 25% higher, with the top earners again seemingly getting a bigger increase. The volatility in the "total earnings" graph after rank 2 000 000 reflects the fact that these individuals generally have a choice as to whether to file, and besides additional benefits, they also record more deductions.

The relationship between "regular earnings" and "taxable income" is shown in Figure 5. This

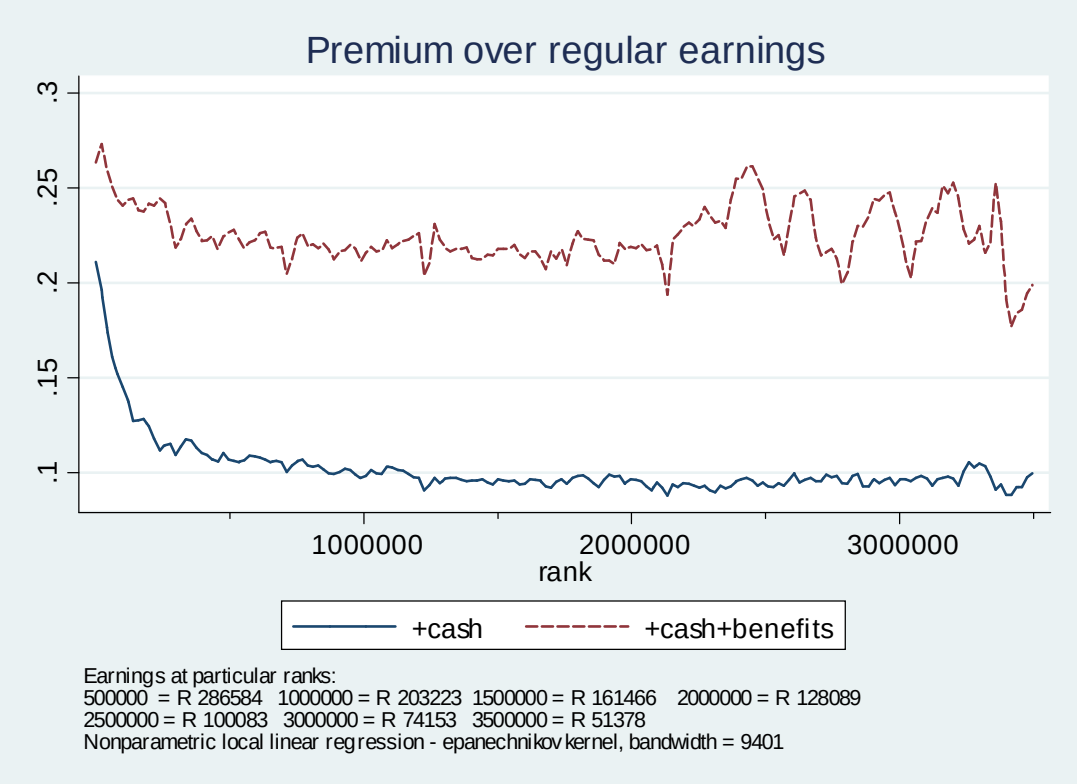


Figure 4: The gap between regular earnings, cash earnings and total earnings in the tax assessments data

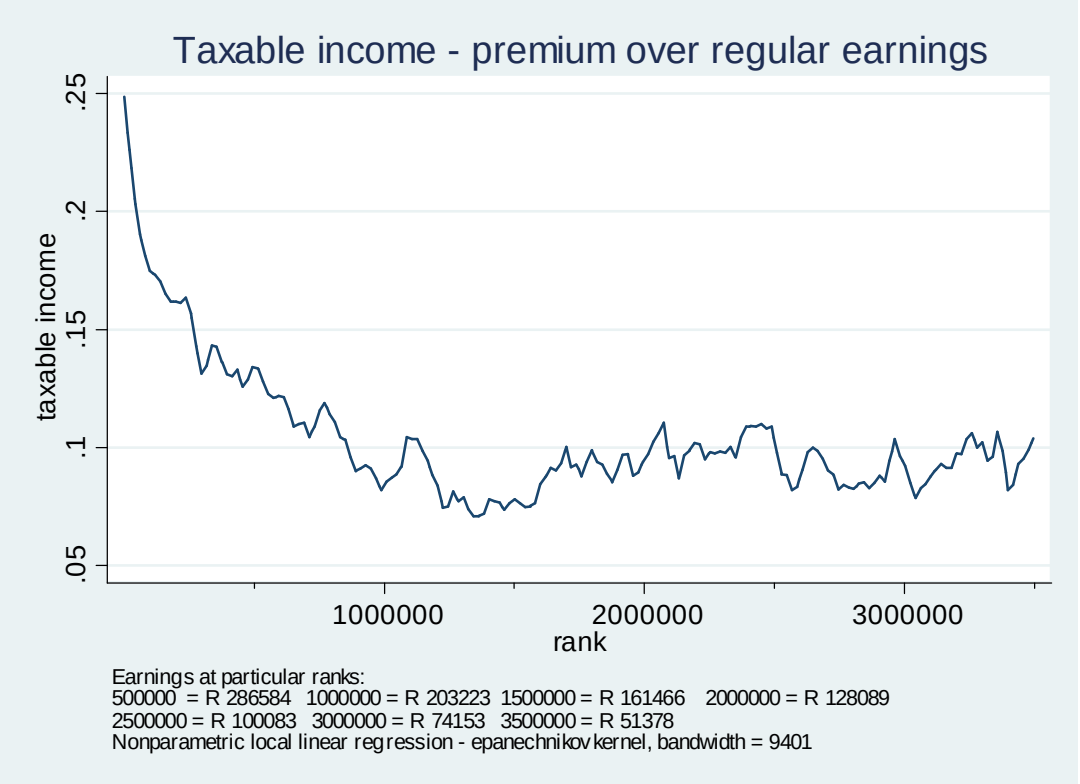


Figure 5: The gap between regular earnings and taxable income for employees

suggests that for most employees taxable income is on average 10% higher than regular earnings, but that for the top earners the gap goes up to 25%. This is a reflection not only of lump sums and benefits as shown in Figure 4, but of investment income and profits (from business sidelines) which some of the highest earners also have.

The graphs are based on average premia in a neighbourhood of particular regular earnings. Table 3 reports the correlation coefficients across the whole distribution between regular earnings (i.e. income3601), the additional earnings due to lump sums and other forms of cash earnings (i.e. cash earnings - regular earnings), the additional earnings due to benefits and allowances (i.e. total earnings - cash earnings) and the difference between taxable income and regular earnings. As is evident all these additional types of earnings are positively correlated with the baseline earnings (some strongly so) suggesting that these additional sources of income all tend to **increase** inequality.

5.2 Employers and own account workers

In the case of the "self-employed", the variable of interest is "gross earnings" as defined in section 3.3.2. Figure 6 shows the relationship between the "tax parade" and the "survey parade" in the case of individuals that we classify as self employed. In this case the two series do not track each other. Instead tax parade starts at a markedly higher level, but then decreases at a much more rapid rate. The two series cross over in the region of rank 54 000, when the earnings are around

	income3601	additional cash	additional benefits	additional taxable
income3601	1.000			
additional cash	0.5508	1.000		
additional benefits	0.4359	0.2799	1.000	
additional taxable	0.5286	0.8166	0.5550	1.000

Table 3: Correlation coefficients between different forms of income for employees

R600 000 (i.e. R50 000 per month). This is well above the compulsory threshold for filing for individuals with regular earnings. Of course individuals with business income are not exempt from filing even if they earn below this.

The gap between the series in Figure 6 is misleading, given the log scale on the y-axis. The size of the mismatch is more evident in Figure 7 where we plot the premia It is evident that at the start of the parade, the tax data suggests earnings that are 100% bigger than those in the survey data, whereas by rank 300 000 the gross earnings reported in the tax data are less than half of the ones in the survey data. Indeed as Figure 6 already indicates, there are many fewer individuals who we identify as "self-employed" in the tax data than there are ones reported as "self-employed" in the QLFS.

6 Discussion

6.1 The quality of the employee earnings' data

The evidence produced above suggests strongly that the gross earnings figures reported in the QLFS look like "regular earnings" (before tax) in the tax data, i.e. they omit annual lump sum payments (e.g. annual bonuses), medical aid payments and other allowances. As a result they are likely to understate full earnings by around 40%, with bigger gaps at the top of the earnings distribution than further down.

The speculation about missing high income earners in the QLFS turns out to be partially true, but not to the extent envisaged. The distribution of employee earnings certainly misses a handful of the highest earners. Below that, however, the QLFS seems to find a lot of high earners, helped in part by what look like curious imputations.

Of course the employee earnings data misses also incomes from sources other than the "main job" (such as rental income, director's income and income from investments). Again this is accrues disproportionately to the top of the earnings distribution.

6.2 The quality of the data for the self-employed

The situation for the self-employed is much worse. It is clear that there is substantial under-reporting or under-coverage of the top income earners among the self-employed. Interestingly, however, the survey data seems to find more individuals with self-employment incomes in the range of several hundreds of thousands a year (but not millions). There are two potential explanations:

- Some of the reported earnings in the QLFS are exaggerated (perhaps to make the individual seem more important)
- There is significant underreporting to the tax authorities of self-employment incomes in this range

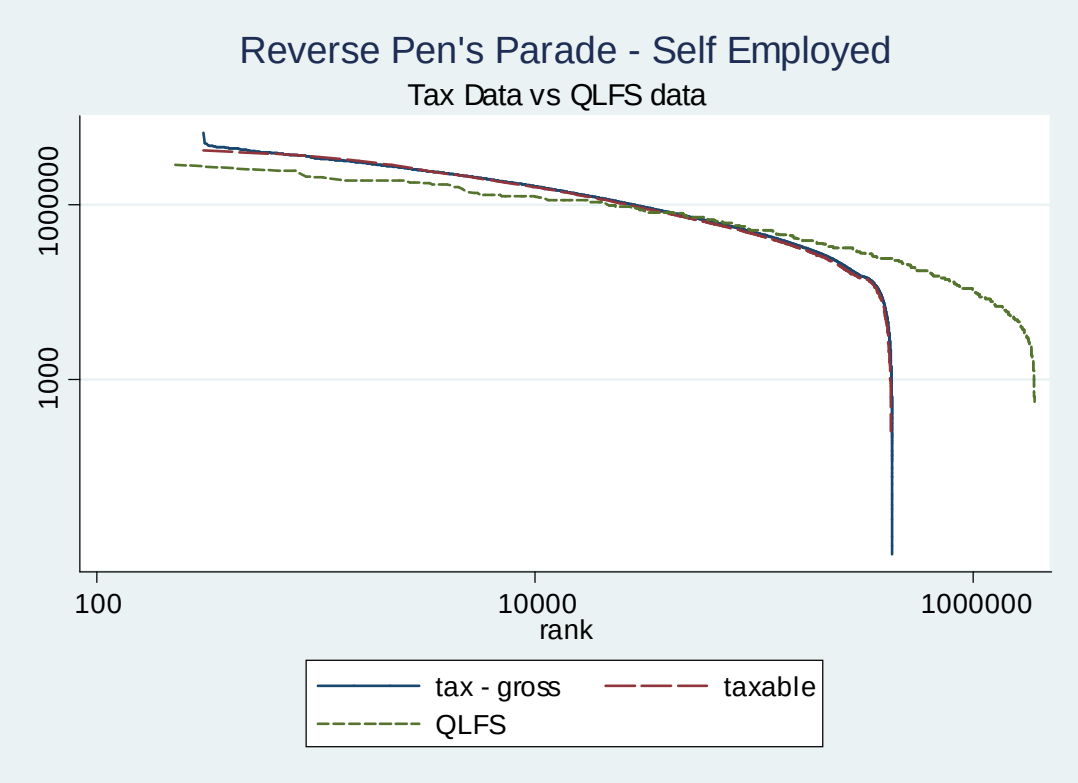


Figure 6: The "tax parade" and the "survey parade" differ noticeably in the case of the self-employed

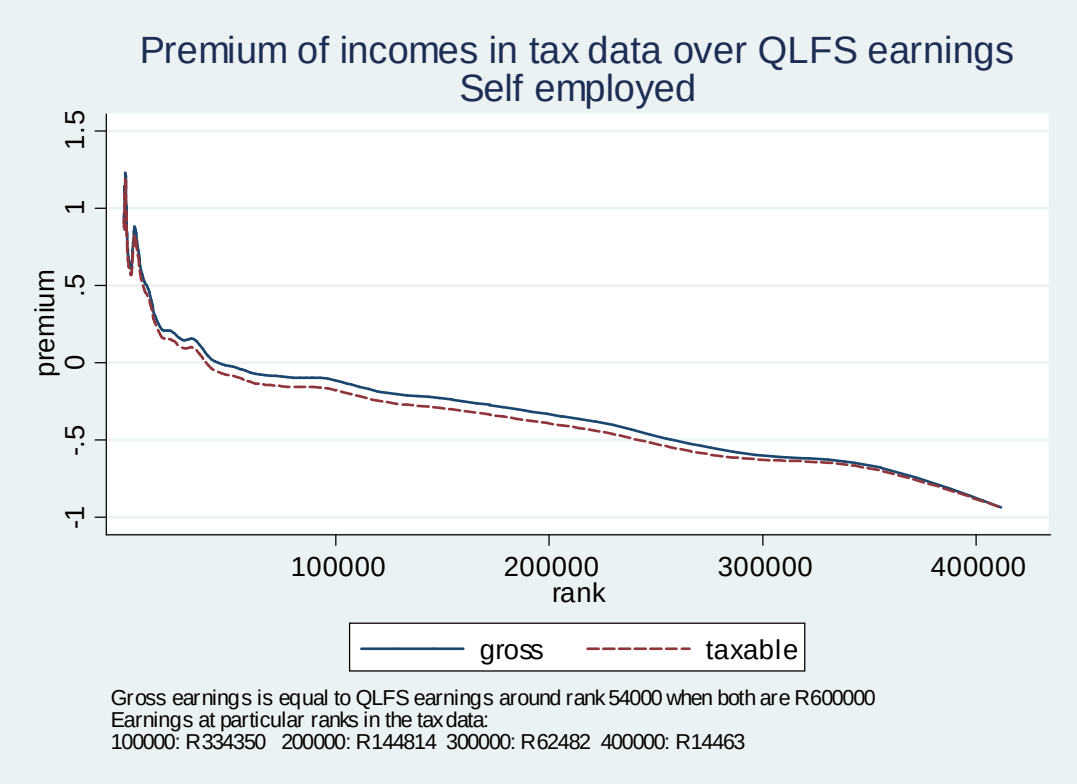


Figure 7: The premium of tax income over survey earnings, i.e. $\frac{(tax-survey)}{survey}$ against rank in the distribution

	Employees	Self-employed	Everyone
QLFS data	0.567	0.684	0.592
Adjusted QLFS data	0.599	0.716	0.620

Table 4: Gini coefficients for earnings adjusting for underreporting

The second hypothesis is very plausible given the fact that these sorts of incomes are unlikely to attract the sort of audit attention of the top earners. Furthermore some of the transactions of skilled mechanics and artisans (who might fall into this income range) may occur in cash and may actually be more difficult to audit.

6.3 Tax data versus survey data

Our findings are in line with the theoretical expectation that tax data would be more accurate and would have better coverage where earnings are easy to audit and reported accurately by third parties (employers in this case), but would be under-declared where they are more difficult to audit and where the likelihood of being monitored is low. This would explain why the earnings data in the surveys tracks the regular earnings in the tax data, but the slopes of the "tax parade" and "survey parade" are very different when looking at individuals whose main source of income is from profit and investment returns.

6.4 The implication for the study of inequality

We saw that the different additional sources of income tended to be correlated with income, thus tending to exacerbate inequality. In order to get some fix on the possible implication, we take the premia that we estimated in Figures 3 and 7 and use them to inflate the earnings at the corresponding positions in the survey distribution, with the following exceptions:

- In the case of the QLFS, we inflate all earnings from wage employment by 20%, except if the premium is predicted to be higher (i.e. at higher positions in the distribution)
- If the predicted premium in the self-employment distribution is negative, we do not adjust the earnings downwards.

The impact of these adjustments on the Gini coefficients for labour earnings is shown in Table 4. We see that the Gini coefficient for labour earnings in the pooled four waves of the QLFS is 0.567, which is very similar to the estimates shown in Wittenberg (2017b, Figure 1, p.299). Once we inflate these earnings upwards in line with the premiums shown in Figure 3 the Gini rises by three points to 0.599. Similarly the Gini coefficient for self-employment income (inflated upwards in line with the premia shown in Figure 7, where positive) also goes up by three percentage points to 0.716. Consequently inequality of overall earnings also rises three percentage points. Unsurprisingly the distribution of self-employment income is considerably more unequal than income from employment.

In order to get a sense of what this type of inequality means, we end this analysis by reporting the traditional Pen's parade for the overall earnings distribution (i.e. both employee and self-employed), using the adjusted QLFS data as our guide and norming the average height to 1.79m, as was done in the parade on personal per capita income cited earlier.

The dwarf leading this parade is not even one centimetre tall. After a minute of the parade (and 216 000 individuals), the person passing by is still only 8 cm tall. By the time a quarter of the parade has passed after fifteen minutes and 3.2 million people, the height of the dwarf

marching past is 38 cm - around shin height. At the half an hour mark the parade has reached double that height i.e 76 cm, which is still below the belt. It is only in the 45th minute that someone of average height comes past. And as in the previous parade, the heights increase rather rapidly. At the 50th minute the height is at 2.91 m, at 55 minutes we're up to 5 m and at the 59th minute an individual 10.9 m tall strides past. In the last four-hundredths of a second the tallest person in the adjusted QLFS earnings parade comes by with a height of 144 m. If we take the tax data we would still see an individual of over 300m before the end.

In broad outlines it is remarkable how similar this parade looks to the income parade quoted earlier. The person with average earnings comes past a few minutes earlier, but in broad outlines the inequality in these adjusted earnings is very similar to overall inequality in 1993. Indeed Wittenberg (2017b) shows that the overall trends in inequality mirror those in earnings very closely – except that overall inequality is always somewhat higher because of the unequal way that labour income is distributed across households. The implication is that inequality measures for overall income inequality should probably be a few percentage points higher, due to the measurement issues suggested by Table 4. It is clear, as most analyses have suggested (Leibbrandt, Woolard, Finn and Argent 2010), that inequality hasn't budged since the end of apartheid.

7 Conclusion

Our analysis suggests that there are measurement and coverage errors in both the survey and the tax data. It seems that the earnings question for employees is eliciting before tax regular earnings, but missing annual bonuses, medical aid contributions and other benefits. The errors are likely to be bigger at the top of the distribution, implying that inequality measures will be underestimated. In the case of self-employment, it appears that the gaps are larger but that the tax data is finding too few incomes for individuals earning just below R50 000 per month. The overall impact of mismeasurement is again likely to understate inequality. Overall we suggest that earnings inequality as measured by the Gini coefficient may be three percentage points higher.

References

Daniels, R. C.: 2012, Questionnaire design and response propensities for employee income micro data, *Working Paper 89*, SALDRU. Revised version. Available from opensaldru.uct.ac.za.

Edwards, L., Sanfilippo, M. and Sundaram, A.: 2016, Importing and firm performance: New evidence from South Africa, *Working Paper 2016/39*, UNU-WIDER.

Finn, A.: 2015, A national minimum wage in the context of the South African labour market, *Working Paper 153*, SALDRU, University of Cape Town.

Isaacs, G.: 2016, Zuma's national minimum wage "own goal", *GroundUp* **15 February 2016**. http://www.groundup.org.za/article/zumas-national-minimum-wage-own-goal.

Kerr, A.: 2016, Job flows, worker flows, and churning in South Africa, *Working Paper 2016/37*, UNU-WIDER.

Kerr, A., Lam, D. and Wittenberg, M.: 2016, Post-Apartheid Labour Market Series [dataset], DataFirst, University of Cape Town. Version 3.1.

Leibbrandt, M., Woolard, I., Finn, A. and Argent, J.: 2010, Trends in South African income distribution and poverty since the fall of Apartheid, *Social, Employment and Migration Working Papers 101*, OECD. http://dx.doi.org/10.1787/5kmms0t7p1ms-en.

Orthofer, A.: 2016, Wealth inequality in South Africa: Evidence from survey and tax data, *Working Paper 15*, REDI3X3.

Pen, J.: 1971, *Income Distribution*, Allen Lane, The Penguin Press, London.

Pieterse, D., Kreuser, C. F. and Gavin, E.: 2016, Introduction to the South African Revenue Service and National Treasury firm-level panel, *Working Paper 2016/42*, UNU-WIDER.

Seekings, J. and Nattrass, N.: 2015a, 'National' minimum wage setting in South Africa, *Working Paper 362*, CSSR, University of Cape Town.

Seekings, J. and Nattrass, N.: 2015b, *Policy, Politics and Poverty in South Africa*, Palgrave Macmillan, Houndmills, Basingstoke, United Kingdom.

Statistics South Africa: 2010, Quarterly Labour Force Survey Questionnaire 2010:Q3.

Tregenna, F. and Tsela, M.: 2012, Inequality in South Africa: The distribution of income, expenditure and earnings, *Development Southern Africa* **29**(1), 35–61.

Wittenberg, M.: 2012, Economics and transformation: Measurement, models, maths and myths, *Inaugural Lecture 8 August 2012*, University of Cape Town.

Wittenberg, M.: 2014, Analysis of employment, real wage, and productivity trends in South Africa since 1994, *Conditions of Work and Employment Series 45*, International Labour Office.

Wittenberg, M.: 2017a, Wages and wage inquality in South Africa 1994-2011: Part 1 – wage measurement and trends, *South African Journal of Economics* **85**(2), 279–297. http://dx.doi.org/10.1111/saje.12148.

Wittenberg, M.: 2017b, Wages and wage inquality in South Africa 1994-2011: Part 2 – inequality measurement and trends, *South African Journal of Economics* **85**(2), 298–318. http://dx.doi.org/10.1111/saje.12147.

Wittenberg, M. and Pirouz, F.: 2013, The measurement of earnings in the post-apartheid period: An overview, *Technical Paper 23*, DataFirst. available at http://www.datafirst.uct.ac.za/images/docs/DataFirst-TP13_23.pdf.

The **Research Project on Employment, Income Distribution and Inclusive Growth (REDI3x3)** is a multi-year collaborative national research initiative. The project seeks to address South Africa's unemployment, inequality and poverty challenges.

It is aimed at deepening understanding of the dynamics of employment, incomes and economic growth trends, in particular by focusing on the interconnections between these three areas.

The project is designed to promote dialogue across disciplines and paradigms and to forge a stronger engagement between research and policy making. By generating an independent, rich and nuanced knowledge base and expert network, it intends to contribute to integrated and consistent policies and development strategies that will address these three critical problem areas effectively.

Collaboration with researchers at universities and research entities and fostering engagement between researchers and policymakers are key objectives of the initiative.

The project is based at SALDRU at the University of Cape Town and supported by the National Treasury.

Tel: (021) 650-5715

www.REDI3x3.org