WIDER Working Paper 2020/62

Earnings in the South African Revenue Service IRP5 data

Andrew Kerr*

May 2020

United Nations University World Institute for Development Economics Research

wider.unu.edu

Abstract: The IRP5 and IT3(a) tax data from the South African Revenue Service have been made available to researchers through a joint project between the South African Revenue Service, the National Treasury, and UNU-WIDER. In this paper, I explain how to use these data to correctly identify labour income for employees, which is crucial for labour market research. I then compare total labour income and employment between 2011 and 2017 with other sources of data, including Statistics South Africa's Quarterly Labour Force Surveys and Quarterly Employment Statistics firm surveys. Finally, I use the data to estimate Gini coefficients and percentile ratios for labour income over this period.

Key words: employment, labour income, South Africa, tax data

JEL classification: C82, D31

Acknowledgements: I thank Amina Ebrahim, Aalia Cassim, and Aroop Chatterjee for helpful comments and suggestions, as well as participants in a UNU-WIDER Metadata workshop in December 2019. The Stata do file that creates the output from this paper is available at the National Treasury Secure Data Facility.

* DataFirst, University of Cape Town, South Africa, andrew.kerr@uct.ac.za

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

ISSN 1798-7237 ISBN 978-92-9256-819-1

https://doi.org/10.35188/UNU-WIDER/2020/819-1

Typescript prepared by Joseph Laredo.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

The IRP5 and IT3(a) tax data from the South African Revenue Service (SARS) have been made available to researchers through a joint project between SARS, the National Treasury (NT), and UNU- WIDER. These data are a new and valuable source of information about several sources of income paid to individuals. One of the main uses of these data is to undertake research on earnings and jobs for individuals who are employed. Unfortunately, most of the researchers undertaking labour market research using the IRP5 data thus far have made what may be an error in separating labour and non-labour income, and thus being able to separate jobs from other income sources, such as pension payments.

In this paper I explain how to use these data to correctly identify labour income for employees, which is crucial for labour market research. Identifying labour income is the only way to identify jobs and employment; incorrectly including non-labour income will mean that individuals who are not employees but who receive non-labour income will be incorrectly assumed to be employed. Identifying labour income for the self-employed is more complicated and would require the use of the IRT12 assessment data. I do not undertake this in the paper, but it is an important issue to resolve.

In the next section I briefly describe the IRP5 data. In Section 3, I explain the aggregate income source codes in the IRP5 data. In Section 4, I explain how one can identify labour income, jobs, and the employed. I compare several methods of identifying labour income, and show how they differ. In Section 5, I compare total income in the IRP5 data with external sources of labour income data, including the National Accounts, Quarterly Labour Force (household) Survey, and Quarterly Employment Statistics, a firm survey. In Sections 6–9, I discuss the measurement of job duration and measurement error, estimate trends in jobs, explain the recommended method of preparing the IRP5 data for labour market research, and give a brief comment on measuring inequality in labour income for employees. Section 10 concludes.

2 Description of the IRP5 data

Pay As You Earn (PAYE) tax-registered companies have been required, since the 2011 tax year, to file IRP5 tax certificates for all employees earning more than ZAR2,000 per year who have tax deducted, and IT3(a) certificates for those who earn more than ZAR2,000 per year who do not have employee tax deducted (Pieterse et al. 2018).[1] IRP5 certificates are also filed by financial services companies for clients receiving other forms of income, such as pension fund or retirement annuity income. Non-natural persons, such as clubs and associations, can also be issued IRP5 certificates. The individual IRP5 and IT3(a) certificates have been made available to researchers as part of a project between SARS, the National Treasury, and UNU-WIDER. I use data from the 2011–2017 tax years, since before 2011 it was not compulsory for firms to issue certificates for all workers.

2.1 Identifying labour income and jobs

[1] In the remainder of the paper I refer to the combined IRP5 and IT3(a) certificate data as the IRP5 data.

In using the IRP5 data for labour market research it is important to distinguish between what is labour income (earnings) and what is not labour income. Income can be defined as the net flow of all payments received in a given period. Labour income, or earnings, is a subset of income—it is income that is obtained specifically from employment. The only way to identify individuals who are employed is to check which of them have IRP5 certificates indicating that they have received income related to employment. This is very different from an analysis of household survey data, where employment is identified first, and then those who have been identified as employed are asked about their earnings.

The IRP5 certificate data contain income reported for individuals from all income sources that require an IRP5 certificate to be issued. There are incomes from some sources that do not appear in the IRP5 data. For example, rental incomes from a property or business income are not included, even if these incomes are reported to SARS and included in an individual's tax calculation. They will appear in the assessment/ITR12 data.

The income reflected on each tax certificate is broken down into disaggregated 'source codes' that identify the various sources of income. For example, an employee could receive income from employment under source code 3601, and the medical aid contributions of the employer would be listed separately under source code 3810. Each certificate lists all the disaggregated income source codes associated with that certificate. The main contribution of this paper is to describe how these disaggregated source codes can be used to identify labour income that resulted from employment, and thus to identify the employed individuals within the IRP5 data.

3 Understanding the aggregate income source codes

Before July 2019, the IRP5 data made available to researchers did not provide the disaggregated source codes just discussed. Instead, the data had three aggregated income variables. These variables were named total_grossretfundincomeamnt (Gross retirement funding income, source code 3697), total_grossnfundincomeamnt (Gross non-retirement funding income, 3698) and total_grossntaxableincomeamnt (Non-taxable income, 3696) in the IRP5 dataset. Most researchers using the IRP5 to understand the labour market were advised at the time that the sum of the amounts under each of these three aggregate codes on an IRP5 certificate was labour income, which I argue is incorrect.

UNU-WIDER and the National Treasury subsequently created a new version of the IRP5 data for use by researchers, called the IRP5 panel (Ebrahim and Axelson 2019), which is actually four different panel data sets. These are called the IDs panel, the Employment Panel, the Source of Income Panel, and the Income Panel. Only the Source of Income Panel includes the detailed source codes described above and this part of the data has been ignored by most researchers. In my own analysis I did not use the IRP5 panel. Instead, I used a version of the IRP5 data with disaggregated source codes that was created by UNU-WIDER in the process of creating a firm-level panel, which has data from the IRP5 eventually collapsed to the firm level.

My argument is that using disaggregated source codes is essential for creating a measure of labour income and identifying jobs. I next explain why using the aggregate codes is incorrect. There are two main pieces of evidence that summing the three aggregate codes to arrive at labour income is incorrect. The first is the National Treasury Explanatory Memorandum on the Taxation Laws Amendment Bill 17b of 2016, which implies that source code 3698 (Gross non-retirement funding income) contains non-labour income. The memorandum states that 'non-retirement funding income […] included passive income such as interest or royalties' (National Treasury 2016).

The second piece of evidence is from Tax Tim, a website designed to help individuals fill in their tax returns. The website (Tax Tim 2014, 2018) explains the three aggregate income source codes using an example in which an individual has obtained ZAR24,000 from a retirement annuity fund (RAF), and this is listed as having the disaggregated source code 3610. In the Tax Tim example this income is listed under Gross non-retirement funding income (source code 3698), implying that this aggregate source code contains some income that is not related to the employment of the person that was issued the certificate.

One explanation for the confusion is that there is some ambiguity in the documentation relating to the codes about whether they include labour earnings. Source codes 3697 and 3698 are described in some SARS documentation as 'Gross retirement funding employment income' and 'Gross non-retirement funding employment income', respectively (e.g. SARS 2018), suggesting that these codes include only labour income. But other documentation excludes the word 'employment' from the descriptions (for example SARS n.d.), suggesting that the income is not only from employment. Perhaps because of this ambiguity, the main data set made available to researchers contained only these three aggregate codes, and whilst one of the four datasets in the IRP5 panel (Ebrahim and Axelson 2019) did contain disaggregated source codes, new users were not directed to these data and were not made aware of the difference between income and labour income. Further data from the IRP5 certificates was, however, available as part of the creation of the firm-level panel, which is what I have used in my analysis, and is what I recommend other users focusing on labour income should use.

3.1 Changes in the 2017 tax year

Above I noted that there were three aggregate source codes. But in the 2017 tax year there were only two aggregate codes. For 2017, source code 3699 is gross employment income (taxable), and the description is 'the sum total of all amounts for all income source codes **NOT** included in code 3696' (SARS 2018) [emphasis as in original]. 3699 is thus the sum of aggregate codes 3697 and 3698. Aggregate code 3696 is described as 'the sum total of all income amounts indicated as non-taxable', suggesting that this code is unchanged in 2017.

4 Using disaggregated source codes to identify labour income

The main conclusion from the discussion above is that researchers have incorrectly used aggregate source codes 3697, 3698, and 3696 (or 3696 and 3699 in 2017) to identify labour income. Because they are the only way to identify jobs and the employed, these aggregate codes have thus also led to incorrect definitions of jobs and the employed. I now explain how the disaggregated source codes available in the IRP5 data can be used to identify labour income, jobs, and employment, building on the work of Pieterse et al. (2018).

Pieterse et al. (2018) provided a list of 22 disaggregated income source codes that they argued were employment-related, and that could be used to identify labour income. They used this list of codes in the construction of a firm-level panel dataset containing SARS Company Income Tax data. However, most researchers undertaking individual or job-level (rather than firm-level) analysis of the labour market using the IRP5 data have not used Pieterse et al.'s list, though some have used the disaggregated source codes. Kerr (2016, 2018) calculated labour income using only one of the 22 disaggregated source codes, 3601, but thus incorrectly excluded some jobs and labour income.

Wittenberg (2017)[2] and Bassier (2019) constructed labour income from a few of the more important employment-related source codes. I use the IRP5 data below to analyse how these measures of labour income compare.

As part of the current research, I reviewed the Pieterse et al. (2018) list of employment-related source codes, and found that there were additional codes that indicated income from employment. The most important are allowances paid to employees (3709, 3710, 3711, 3712, 3713). There are also two codes in the Pieterse et al. (2018) list that I do not think qualify as income for employees, and I would recommend excluding them. These are 3615, Director's remuneration, and 3616, Independent contractors' income. Directors are not necessarily employees of a company[3], and independent contractors are by definition not employees.[4] The new definition of labour income[5] that I am proposing increases the total labour income by an average of 5 per cent relative to the Pieterse et al. (2018) method but increases the number of jobs by less than 1 per cent. Table A1 in the Appendix provides the income source codes used by Pieterse et al. (2018) and myself that are assumed to indicate labour income.

In the following sections I analyse the IRP5 data and compare the various definitions of labour income data that have been used with the one I am proposing. In feedback on my list of labour income codes it has been suggested that some should not be considered labour income. These include codes that are reimbursements for employee expenses (3805, 3804). The total income in these codes is very small, so I would not expect this to make much of a difference to my results. But I would advise other researchers using the list to do their own investigation, especially of the extra codes compared with Pieterse et al. (2018) that have fairly large total incomes or that contribute to extra jobs.

5 Labour income in the IRP5 data

To undertake the analysis of IRP5 data I use version 0.6 of the data (National Treasury and UNU-WIDER 2019), which was made available at the National Treasury secure data facility in the second half of 2019. Figure 1 shows total labour income from five methods of calculating labour income in the IRP5 data. The dark green dots show the incorrect method: assuming that the aggregate codes 3698, 3697, and 3696 (or 3696 and 3699 in 2017) constitute labour income. The red dots show the Pieterse et al. (2018) method: using 22 employment-related source codes. The dark blue dots show the method I used, which adds what I believe are several extra employment-related source codes, to get what I think is the most accurate definition of labour income. I also include the Kerr (2018) and Wittenberg (2017) methods for comparison.

Figure 1 shows that incorrectly using all income in the three aggregate codes results in total yearly income across all certificates that is 5–8 per cent higher than the yearly labour income obtained from the disaggregated source codes I am suggesting correctly identify labour income. It also

[2] Wittenberg (2017) used a different data set, the SARS assessment sample from the 2011 tax year, but was able to break down income by the same source codes that are in the IRP5 data.

[3] https://dommisseattorneys.co.za/blog/directors-also-employees/

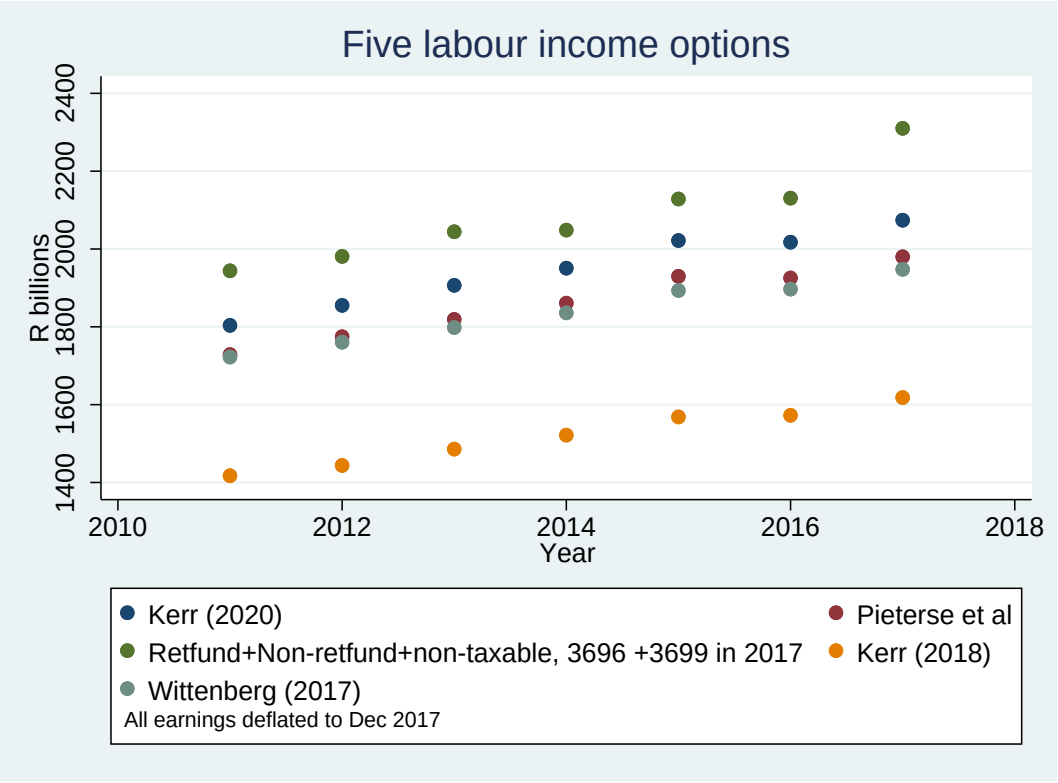
[4] https://www.sars.gov.za/AllDocs/LegalDoclib/Notes/LAPD-IntR-IN-2012-17%20-%20Employees%20Tax%20Independent%20Contractors.pdf

[5] The additional disaggregated source codes I include as identifying labour income are 3702, 3704, 3708, 3713, 3714, 3715, 3751, 3752, 3753, 3757, 3763, 3764, 3765, 3768, 3805, 3810, 3813, 3814, 3815, 3816, 3856, 3858, 3860, 3865, 3803, 3804, 3852, 3855 and 3863. Some of these appear very rarely in the data.

shows that the Pieterse et al. (2018) method of obtaining labour income from 22 source codes results in total labour income that is around 5 per cent lower than the method I am suggesting, using a number of additional codes. The Kerr (2018) method used only the 3601 source code (albeit the one with the largest fraction of labour income), and so substantially understates total true labour income. The Wittenberg (2017) method, using a few important source codes, looks very similar to the Pieterse et al. (2018) method.

Figure 1 also shows that the difference between the summation of the three aggregate codes and my preferred labour income calculation method is quite a bit larger in 2017 than in the other years; the difference is around 11 per cent. It is not clear why there is a jump in 2017. It may be related to the change from three aggregate codes to two.

Figure 1: Comparison of various measures of labour income



Notes: aggregate code 3697 is Gross retirement fund income, 3698 is Gross non-retirement fund income, 3696 is Non-taxable income and 3699, which appeared for the first time in 2017, is Gross renumeration.

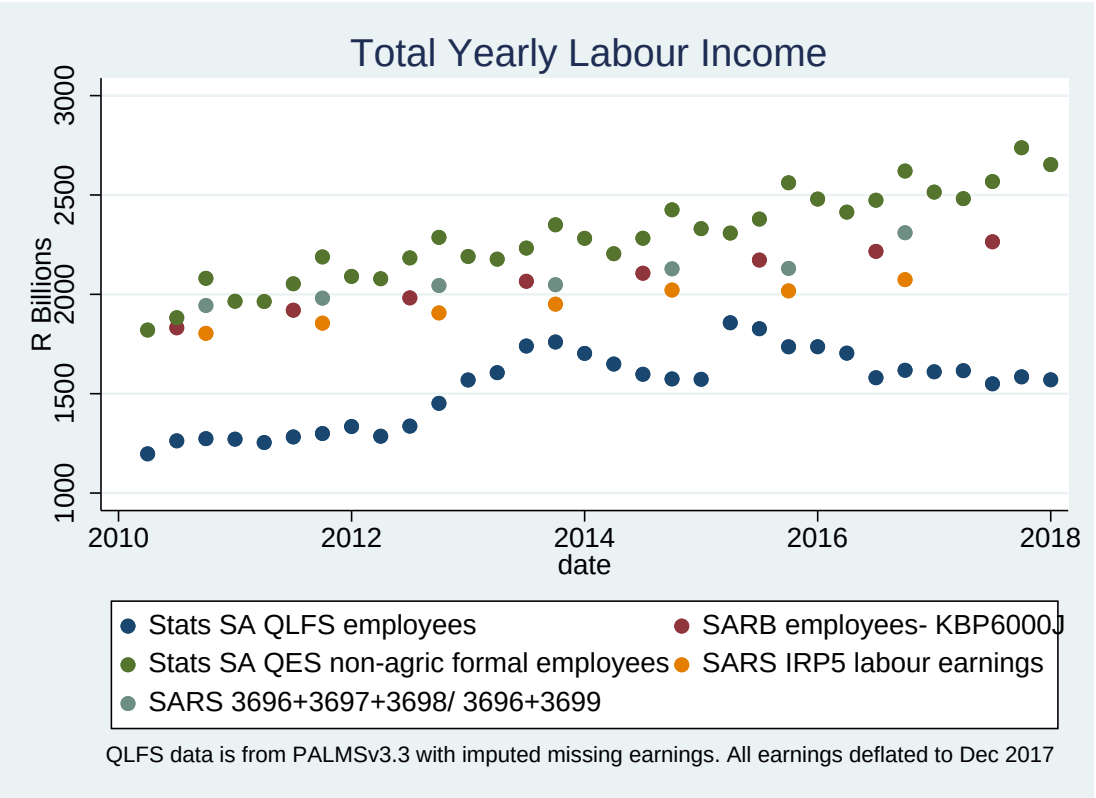
Source: author's illustration.

5.1 Comparisons of alternative labour income data sources

Having shown that summing the income in the three aggregate codes results in a measure of income that is too large, and that includes non-labour income, we need to check whether the various definitions of labour income result in totals that are similar to those obtained from non-SARS sources of labour income data. Figure 2 shows total real labour income from the IRP5 data, the Stats SA Quarterly Labour Force Survey (QLFS) household survey data, the Stats SA QES firm survey data, and the South African Reserve Bank total employee compensation estimate. Table 1 shows the ratios of the last three labour income sources to the SARS IRP5 labour income total in each year.

Figure 2 and Table 1 illustrate a number of important issues. The IRP5 labour income data seem to give a reasonable estimate of total labour income when compared with the other estimates, which is reassuring. The National Accounts estimates of employee compensation obtained from the South African Reserve Bank (SARB) are 4–8 per cent higher than the IRP5 totals. The estimates of gross earnings of employees from the Stats SA QES firm survey are substantially higher than the IRP5 totals (and the QES does not include any income from formal sector agriculture, with which the ratio to the IRP5 total would be even higher). The estimated QLFS household survey totals are around 80 per cent of the IRP5 totals.

Figure 2: Comparison of a variety of sources of labour income



Notes: these data are the revised data, and so do not exactly match the earnings totals from the QES release documents; see Stats SA (2019b). QLFS and QES are estimates from a sample. I have not included the standard errors, but they are very small.

Source: author's illustration from IRP5, QES aggregate, PALMS, and SARB data. The source of the QES data is Stats SA (2019a).

Table 1: Ratio of 3 other labour income sources to total IRP5 labour income

Year	QLFS household surveys	QES firm surveys	SARB KBP6000J
2011	0.71	1.15	1.01
2012	0.70	1.18	1.02
2013	0.76	1.20	1.02
2014	0.90	1.20	1.04
2015	0.78	1.20	1.03
2016	0.86	1.27	1.07
2017	0.78	1.26	1.06

Notes: the ratios are calculated from the data shown in Figure 2, using Quarter 2 data each year for QLFS household surveys and QES firm surveys.

Source: author's calculations.

It is beyond the scope of this paper to provide a detailed interrogation of the differences between the various sources of labour income data, but it is worth making a few comments. The first is that, as Figure 2 and Table 1 clearly illustrate, the IRP5 data and firm survey data capture much more labour income than the QLFS.[6] This is to be expected, since there is item and unit non-response in the QLFS, as well as likely underreporting of earnings by those who respond to the earnings questions.[7] This result accords with the findings of Kerr and Wittenberg (2017), who undertook a similar exercise and showed that the labour-related household surveys had recorded around 80 per cent of the employee compensation total as recorded in the National Accounts (series KBP6000J) between 1993 and 2014.

Figure 2 also shows that real total labour income for employees in the QLFS was declining from 2013—apart from a jump in 2015. The declining real total labour income in the QLFS is surprising. The jump in 2015 is likely a result of the introduction of a new 'master sample' of enumeration areas by Statistics South Africa in the first quarter of the 2015 QLFS. This new master sample was based on the 2011 census, so in theory it provides a more accurate picture of the South African labour market than the prior master sample, which was based on the 2001 census. Kerr and Wittenberg (2019) note that the change in master sample in 2015 has impacted estimates of household size, unemployment, and employment. Figure 2 shows that the change also seems to have affected the QLFS estimates of total labour income in a positive manner—bringing them closer to the other sources of data, although the subsequent declines are concerning.

Kerr and Wittenberg (2019) also discuss the imputations in the QLFS undertaken by Stats SA and the massive variation across waves of the Gini coefficient estimated using the QLFS, which the authors hypothesize is partly the result of these imputations. The declining real total labour income in the QLFS shown above is another serious concern about the QLFS earnings data, but further investigation is beyond the scope of this paper. As Kerr and Wittenberg (2019) have argued, obtaining unimputed QLFS earnings data from Stats SA would be very useful in determining whether the imputations are responsible for these worrying trends.

A final concern is that the QES total labour income estimate for employees is around 20 per cent higher than the IRP5 and National Accounts estimates of total labour income, despite its excluding the earnings of those in agriculture and all workers in non-VAT registered (i.e. informal) firms. Further investigation is required to understand this discrepancy.

---

[6] The estimated total labour income in the QLFS excludes the self-employed, whose earnings are also not captured in the IRP5. However, the QLFS estimate includes all employees, including those who work in non-PAYE registered firms (for example, informal, unregistered firms). It is difficult to use the QLFS to isolate those who are in formal firms, so I have included the labour income of all employees. Since informal workers do not appear in the SARS data, the ratio of QLFS to IRP5 labour income is higher than it would be if these workers were excluded from the QLFS estimates shown in Figure 2.

[7] It should be noted that in the QLFS, Stats SA imputed for those with both bracket responses and complete refusals to earnings questions up until 2012 Q2. After Q2 2012 those with bracket responses were still imputed, but those who provided no information on their labour income were not imputed. The data shown in Figure 2 include further imputations undertaken for PALMS (Kerr et al. 2019), which need to be included to make sensible comparisons about total labour income with other datasets. One potential explanation for the trends shown in Figure 2 is that the imputations undertaken for PALMS by Kerr et al. (2019) are responsible for the constant or declining estimates of total labour income. However, total labour income when using only the Stats SA imputed data displays similar trends (these results are not reported here, but are available from the author).

6 Accounting for job duration

The labour income that appears on the IRP5 certificate is for the entire tax year. Most researchers would want to create a labour income measure adjusted for job duration or hours worked during the year for each job. Unfortunately, there are no sources of data on hours worked. However, there are two sources in the IRP5 certificate from which job duration can be obtained. The first is the start and end dates of employment, which have been used to construct monthly labour income by a number of researchers, including Bassier (2019), Bhorat et al. (2017), Cassim and Casale (2018), and Kerr (2018) (and the earlier working paper version, Kerr (2016)).

Kerr (2018) showed that the start and end date data implied trends in employment within each tax year that did not seem plausible, and suggested that this might be due to HR administrators having little incentive to fill out this part of the certificate correctly. If employment duration is measured with error, monthly earnings created from start and end date information will also be measured with error. This has not been investigated or discussed by researchers using the tax data subsequent to Kerr (2018).

The second source of data on job duration is the periods of assessment and periods worked recorded on the IRP5 certificate. This gives the units in which work time is recorded (for example, days, weeks, or months) and the number of those units that the individual worked. Both sources of duration data were used to construct the monthly earnings incorporated into the firm-level panel, as described by Pieterse et al. (2018), but the work on individual labour income does not seem to have used this alternative source of data to calculate labour income per unit of time worked. The weakness of both methods is that one cannot tell if an individual was working 1 hour per week or 40 hours per week, only that (for example) an individual was employed for 1 month in the tax year.

Given that there are two ways to calculate job duration, and that creating a monthly labour income is important, it is worth comparing the two sources of job duration data. The last column of Table 2 shows that around 10 per cent of certificates have impossible/incorrect combinations of start and end dates—for example, start dates in the previous tax year or end dates in the subsequent tax year. However, it turns out that most of these 'impossible' start and end date certificates are certificates that start in the last two weeks of February, i.e. the final two weeks of the following tax year, and most of them then end just before these dates in the following February. Many government employees are paid on the 15th of the month, which may be part of the explanation. It thus seems that these certificates should be treated as legitimate, rather than discarded or be adjusted.

The second source of job duration information appears to have less obvious measurement error, shown by the much smaller fraction of missing or impossible periods worked and periods assessed. There are some observations where periods worked are listed as 0.1 or 0.2; these are likely to be 1 or 2, but the error is much smaller. There are almost no observations in which the number of units of time worked exceeds the maximum possible number of units—e.g. no one reports 52 periods worked (likely to be 52 weeks worked, i.e. the full year) when the periods of assessment in the year are listed as 12 (i.e. months). This may because the IRP5 forms do not allow such combinations to be entered.

Table 2 also shows several percentiles and the mean of the duration distributions for the two methods of calculating job duration. They look fairly similar, which is reassuring. The median job duration is 1, i.e. the median certificate was for the entire year, and the mean is between 0.71 and 0.73.

Table 2: Fraction of year worked statistics

Year	Method	25th percentile	Median	Mean	Share of 'impossible' certificates
2011	Periods	0.42	1	0.736	1.3%
	S+E	0.47	0.997	0.745	9.7%
2012	Periods	0.33	1	0.7	1.3%
	S+E	0.41	0.989	0.72	12.1%
2013	Periods	0.35	1	0.71	1.3%
	S+E	0.41	0.989	0.733	12.8%
2014	Periods	0.35	1	0.71	1.3%
	S+E	0.4	0.989	0.72	12.8%
2015	Periods	0.33	1	0.7	1.3%
	S+E	0.35	0.98	0.71	12.2%
2016	Periods	0.36	1	0.714	1.3%
	S+E	0.41	0.989	0.725	10.8%
2017	Periods	0.33	1	0.704	1.3%
	S+E	0.36	0.986	0.711	11.9%

Note: analysis undertaken at certificate level.

Source: author's calculations.

Table 3 shows the correlation coefficients of the two variables using three methods of treating duration derived from the start and end dates and keeping the periods worked and periods assessed duration as reported on the certificate. The first way of treating the start and end dates is to use them as is. This results in correlation coefficients of between 0.72 and 0.82, which seem fairly low, given that the two are measuring the same thing, and are filled in by the same person. The second way of treating the start and end dates is to keep as is all certificates with a start date in the month before the start of the tax year, or an end date in the month after the end of the tax year, if the job duration is less than 366 days. But I then exclude certificates with start dates more than a month before the start of the tax year, and certificates with an end date more than a month after the end of the tax year, since these are likely to be errors. This treatment excludes less than 1 per cent of the certificates and results in much higher correlation coefficients, which is reassuring.

The final method, shown in column 3, is to implement the fix that some researchers have used. This fix is to set the start date to the start of the tax year if it is before the start of the tax year, and to set the end date to the end of the tax year if it is after the end of the tax year. This method produces a very similar correlation coefficient to the second method, although the correlation looks worse in the 2014 and 2015 tax years.

The analysis shown in Table 3 excludes certificates where total income is less than ZAR2,000 per annum, which is the threshold above which firms are required to issue certificates. Table 4 shows the same correlation coefficients as Table 3, but includes certificates of less than ZAR2,000 per month. Clearly, the correlation between the two ways of measuring job duration is much worse when using the start and end dates as is to create duration. Including those certificates with start dates in February of the previous tax year or end dates in March of the following tax year but excluding other certificates, as shown in column 2, looks better, but the correlation when adjusting these to the first and last days of the tax year, respectively, as shown in column 3, also looks poor.

This suggests that analysts should be wary of using very low income certificates, and my recommendation would be to exclude them entirely from analysis. The final column of Table 4 shows that this would involve excluding an average of 8 per cent of certificates per year that have labour income.

Table 3: Correlation coefficients between two duration measures

Year	As is	<Feb, >mar, <0, >366 =missing	S= 1 Mar, E=28 Feb, dur=365 if >366
2011	0.72	0.94	0.94
2012	0.77	0.93	0.92
2013	0.82	0.95	0.94
2014	0.82	0.95	0.88
2015	0.81	0.95	0.89
2016	0.80	0.95	0.95
2017	0.82	0.96	0.95

Note: each column shows a different method of treating start and end dates.

Source: author's calculations from IRP5 data. Certificates with less than ZAR2,000 pa excluded.

Table 4: Correlation coefficients between two duration measures not dropping <ZAR2000 p/a certificates

Year	As is	S= 1 Mar, E=28 Feb, dur=365 if >365	<Feb, >mar, <0, >366 =missing	% of Kerr method jobs with >0 & <2000 earnings per year
2011	0.144	0.662	0.795	9.5%
2012	0.278	0.686	0.808	9.8%
2013	0.313	0.576	0.815	7.9%
2014	0.327	0.545	0.861	8.1%
2015	0.357	0.597	0.859	9.2%
2016	0.312	0.625	0.862	5.9%
2017	0.372	0.531	0.881	7.4%

Source: author's calculations from IRP5 data. Certificates with less than ZAR2,000 pa included.

7 Jobs estimates

In the discussion above I noted that it is important to correctly identify certificates with labour income because identifying certificates that are jobs is only possible if one correctly identifies labour income. Table 5 shows the fraction of certificates that have positive income in any of the three aggregate source codes that are not actually employment related, using both the Pieterse et al. (2018) method and my own method, which adds several more source codes that I consider labour income. Around 6–7 per cent of 'jobs' identified when using the incorrect aggregate source code definition are actually not jobs associated with labour income. There is only a small difference between the Pieterse et al. (2018) and my method, and the figures in the table suggest that the total number of jobs is only about 1 per cent lower using the Pieterse et al. (2018) method. As discussed above, the difference in aggregate labour income is more substantial, around 5–8 per cent.

In the previous section I noted that the start and end date data are sometimes measured with error, but that some of the 'error' is jobs that start in the month before the start of the tax year, and/or end in the month after the end of the tax year. In Kerr (2018) I used the start and end date data to create a jobs-per-day graph, but I incorrectly excluded job-days before or after the end of the tax year. This then led, incorrectly, to very large drops in total employment in the last two weeks of each tax year.

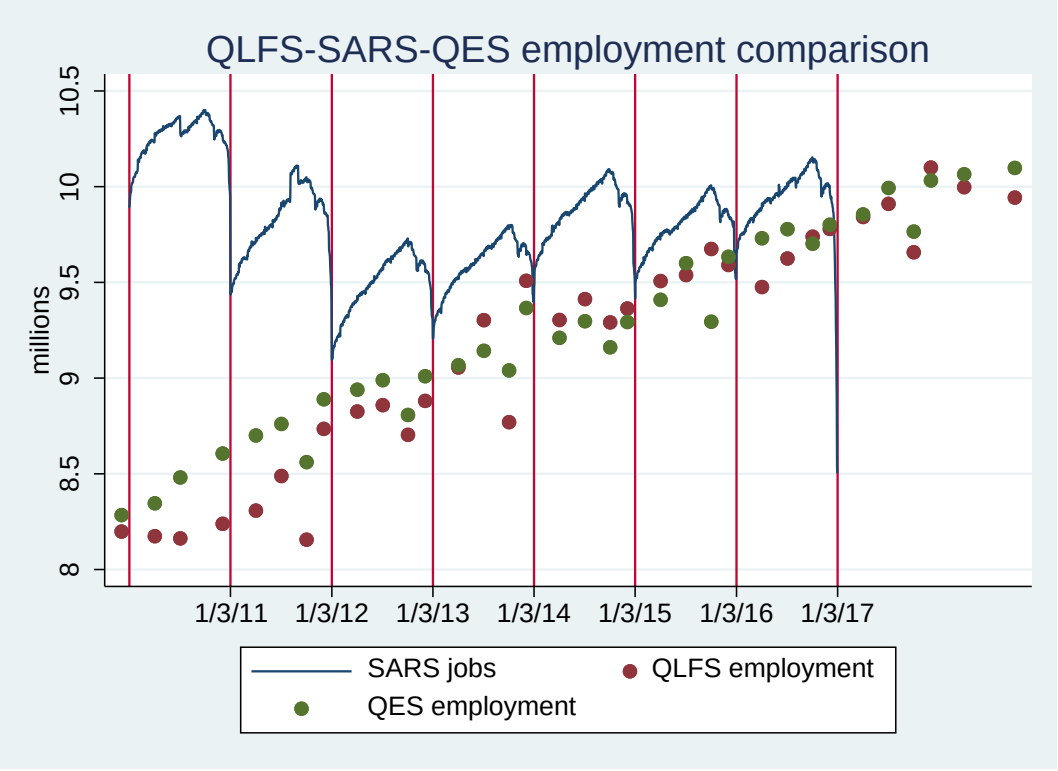
Table 5: Share of 'jobs' identified through 3696/7/8/9 definition that are not employment-related

Year	Pieterse et al.	Kerr
2011	7.7%	7.0%
2012	7.7%	6.8%
2013	10.8%	9.9%
2014	6.6%	5.5%
2015	6.6%	5.5%
2016	7.0%	5.7%
2017	6.5%	5.3%
Average	7.6%	6.5%

Source: author's calculations.

Figure 3 shows an updated version of the figure from Kerr (2018), including jobs with start dates in the month of the previous tax year as jobs on those days in the previous tax year which their tax certificate said were in the previous tax year. This correction means that the huge drop-offs in the number of jobs in the last two weeks of each tax year shown in Kerr (2018) do not occur. Despite this correction, there is still a (much smaller) drop at the end of each tax year/start of the following one, and there is also a peak in December, followed by a decline and then a big drop in the last two weeks of the tax year at the end of February.[8]

Figure 3: Formal sector jobs estimates from QLFS, SARS, and QES



Notes: QES excludes formal sector agriculture. QLFS and QES are estimates from a sample. I have not included the standard errors, but they are very small.

Source: author's illustration based on PALMS (QLFS), QES (source of data is Stats SA, 2019a), and IRP5 data.

[8] I have excluded certificates of less than ZAR2,000 per annum from Figure 3, which I did not do in Kerr (2018).

The QLFS and QES employment estimates can serve as a check on two aspects of the SARS jobs numbers and are also shown in Figure 3. The first check is on is the odd patterns within each year, which are not replicated in the QLFS or the QES (though each has only four estimates in each year). The second check is on the level of jobs. The QLFS does not ask questions that allow one to restrict the sample only to those in PAYE-registered firms who were issued IRP5 certificates. As an approximation I included only individuals with written contracts, who work for an employer, who reported either that their employer was paying Unemployment Insurance Fund (UIF) contributions or that they were employed in government (where no UIF is paid), and who were not domestic workers. The QES sample frame is VAT-registered firms but excludes agriculture.

The three sets of employment numbers seem fairly similar after the 2011 and 2012 tax years. The IRP5 data include anyone earning more than ZAR2,000 per year, and some of the low earners above this threshold would be excluded from the QES and QLFS. The QES also excludes formal sector agriculture, which the QLFS suggests represents around 4 per cent of total formal employment. That the 2011 and 2012 SARS jobs do not look sensible is the result of the conflation, by SARS, of pension and labour income in the source code 3601 in the 2010, 2011, and 2012 tax years. This means that pension-related certificates cannot be excluded in these years directly. Kerr (2018) excluded around 1.1 million and 1.3 million certificates issued by 23 firms in 2011 and 2012, respectively, that seemed to be for income from pension funds, and also noted that there were likely to be around 500,000 certificates relating to pension income that were still included. Pieterse et al. (2018) briefly discuss this issue but do not make any adjustment to the firm-level panel.

No further work has been done to identify the pension income certificates or to allow researchers to exclude them in these years. This may be partly because, subsequent to the working paper version of Kerr (2018), the age of each individual issued with a certificate was included in the data made available to researchers. This means that the vast majority of certificates with probable pension income incorrectly included in code 3601 in 2010–2012 can be excluded by limiting the sample to individuals younger than 65, as has been done by some researchers. I did not do this in creating Figure 3, so researchers should check how much an age restriction solves the over-estimate in the 2011 and 2012 tax years.

8 Recommended treatment of the IRP5 data

The previous discussion has shown that the choice of labour income measure matters for both labour income and number of jobs. There are a number of other considerations for analysts that have been discussed. In this section I briefly suggest an approach to using the data.

The main conclusion of my analysis is that researchers should not use the sum of the two or three aggregate source code variables as labour income. Rather, the disaggregated codes should be used to create labour income. There is not a substantial difference in total income—between 5 and 8 per cent—when using the list of codes that identify labour income that I have proposed and the list provided by Pieterse et al. (2018) (see Appendix Table A1). The difference in total jobs is smaller, around 1 per cent. The main message is that the aggregate source codes should not be used, rather than that the method I have proposed is better than the Pieterse et al. (2018) method.

Either start and end dates, or periods worked and periods of assessment can be used to calculate job duration, with some caveats. The correlation between variables created from the two approaches is fairly high. However, the jobs-per-day figures shown in Figure 3 suggest that there is still some concern about start and end dates, and this deserves careful consideration by users of the data.

There are several other decisions that I would recommend but have not discussed or only briefly mentioned. I recommend dropping certificates where the nature of person variable is not either 'missing' or 'individual'. The variable has a high fraction of missing in 2011 and 2012, but almost none after that. I also recommend dropping duplicate copies of certificates (where the certificate number is the same) and dropping certificates with less than ZAR2,000 per year. There are also a few individuals with extremely large income amounts. In the analysis in this paper I have excluded any individual with labour income of more than ZAR500 million.

I would also recommend that analysts working on the IRP5 data create smaller samples and run code and obtain results using these first, and then use the full datasets once they have fixed up errors. This speeded up my progress dramatically compared with previous work I had done on the IRP5 data.

9 Measuring inequality in labour income for employees

The preceding discussion has shown that important decisions are required before researchers use the IRP5 data to better understand the South African labour market, and that the method of estimating labour income that has most commonly been used is incorrect. In this section I briefly show the implications of a few of the possible choices researchers face when measuring inequality in labour income with the IRP5 data. Inequality is measured in gross labour income, not accounting for tax deductions.

Tables 6 and 7 show stable Gini coefficients and percentile ratios over the period. Table 6 shows 6 Gini coefficients for each tax year. The first two columns use annual data, i.e. they do not adjust for job durations. Of these two, the first uses my proposed definition of labour income and jobs, whilst the second uses the incorrect income measure summing the income in the 2/3 aggregate source codes. The difference is not large, with the incorrect measure slightly over-estimating the Gini by about 1 point. The next two columns show Ginis calculated using the periods of assessment and periods worked data to calculate job duration and then monthly income. Column 3 uses the labour income definition I have suggested, whilst column 4 uses the incorrect method. The difference between the correct and incorrect methods is slightly larger, around 2 Gini points. Finally, columns 5 and 6 calculate monthly labour income Ginis using start and end dates to obtain job duration, keeping those start and end dates in February of the prior tax year and in March of the following tax year and excluding other impossible start and end dates. Here the differences between the correct and incorrect labour income definitions are much more substantial, with the incorrect method generating impossibly large Gini coefficients. This appears to be the result of certificates with non-labour income and very short durations, many of which are just one day.

Table 6: Gini coefficients for several measures of annual and monthly earnings

Year	Annual labour income	Wrong annual labour income	Monthly periods worked	Wrong monthly periods worked	Monthly start and end date	Wrong monthly start and end date
2011	0.63	0.64	0.61	0.63	0.68	0.94
2012	0.64	0.65	0.62	0.64	0.67	0.93
2013	0.63	0.65	0.61	0.63	0.64	0.94
2014	0.63	0.64	0.60	0.62	0.63	0.91
2015	0.64	0.65	0.60	0.62	0.66	0.92
2016	0.64	0.65	0.60	0.62	0.64	0.91
2017	0.64	0.65	0.60	0.62	0.64	0.90

Source: author's calculations.

Table 7 shows 3 percentile ratios using 3 different income measures for 2011 and 2017. Monthly labour income uses periods worked and periods of assessment whilst monthly labour income 2 uses the start and end date on the tax certificate. The table illustrates that inequality trends vary depending on whether one considers annual or monthly income. The p90/p10 ratio goes up when using annual income but down when using monthly income. The p90/p50 ratios move in the same direction between 2011 and 2017, but the p10/p50 ratio decreases using annual income, and increases using monthly income.

Table 7: Percentile ratios in annual and monthly income

Year	Ratio	Annual labour income	Monthly labour income	Monthly labour income 2
	p90/p10	43.8	18.5	20.3
2011	p90/p50	5.7	4.4	4.5
	p10/p50	0.13	0.24	0.22
	p90/p10	52.9	15.9	16.5
2017	p90/p50	6.0	4.7	4.7
	p10/p50	0.11	0.29	0.29

Note: cut-off used is ZAR2,000 per annum.

Source: author's calculations based on IRP5 data.

Given the large number of moving parts in the IRP5 data, it is difficult to make strong conclusions about trends over time, but the data suggest that the Gini coefficients were fairly stable over the period. Comparisons with the household survey data are possible using the monthly labour income measures from Tables 6 and 7, but one should be aware that the QLFS is a point-in-time measure, whereas the IRP5 data presented here are for the whole year.

The IRP5 data include all those that are employed at any time in the year, whereas the QLFS includes only those employed in the reference period. Low-earning individuals are more likely to have short-term jobs (Kerr 2018), so including all those employed any time in the year will lengthen the left tail of the IRP5 earnings distribution, likely increasing inequality relative to a point-in-time measure. One can calculate point-in-time measures in the IRP5 data that are comparable to the QLFS by using the start and end dates, but that is beyond the scope of this paper, and would need to take into account the measurement error discussed above.

10 Conclusion

The IRP5 is an enormously useful source of administrative data. However, in this paper I have documented that using these data for labour market research is not simple. Most of the labour market research undertaken thus far on the IRP5 data at the individual or certificate/job level has incorrectly used an income measure that is a combination of labour and non-labour income. Identifying labour income is important both because labour market analysts would not want to count non-labour income as income from work, and because correctly identifying labour income is the only way to identify IRP5 certificates that are related to work, rather than pensions or other sources of non-labour income.

I have argued that the sum of the 2/3 aggregate codes that has been used is incorrect, and that researchers should use the disaggregated source codes to identify labour income, and then jobs. This has been the approach taken by Pieterse et al. (2018) in creating a firm-level panel, but this approach has not been adopted by most researchers working on labour market issues at the

individual or certificate level, mainly because the versions of the data that researchers were directed to use included only these aggregate codes.

I have suggested a broader set of source codes that identify labour-related certificates than the Pieterse et al. (2018) list (Appendix). This broader set increases the number of work-related certificates by only 1 per cent, but raises total labour income by around 5 per cent, relative to the Pieterse et al. (2018) definition of labour income. Using the 2/3 aggregate source codes to (incorrectly) identify work-related certificates results in an extra 7–9 per cent certificates relative to my method and overstates income by around 5–8 per cent. These are fairly large discrepancies.

The correct definition of labour income results in measures that are comparable to the National Accounts estimates but higher than the Stats SA QLFS household surveys and (surprisingly) around 20 per cent lower than the Stats SA QES firm survey estimates. That the labour income estimates are 25–30 per cent higher than the QLFS estimates confirms that imputation, measurement error, and item and unit non-response are concerns in the QLFS household surveys, as discussed in Kerr and Wittenberg (2019).

One of the main benefits of administrative data like the IRP5 is that they are supposedly free from the sorts of errors in household surveys discussed above, but I have shown this not to be the case. Labour market analysts want to measure job duration to get a monthly labour income number, but the two sources of information both show signs of measurement error. The number of jobs recorded per day within each tax year follows patterns that are not replicated in the QLFS, although the average levels in each year are fairly similar after 2013. The much higher IRP5 job numbers in 2011–2012 can be explained by the conflation by SARS of pension and labour income in the disaggregated source code 3601 in these two tax years. This can likely be solved by limiting the age of individuals in the IRP5 data to 16–64, although non-South African individuals will be missing age data, since these are obtained from the South African ID number.

I have provided a number of recommendations for those wanting to use the IRP5 data for labour market analysis. These include using the disaggregated source codes to construct labour income, dropping duplicate copies of certificates, and dropping certificates where income is less than the ZAR2,000 threshold for mandatory issuing of certificates. I also recommend paying careful attention to the job duration data that can be obtained from the start and end dates or the periods worked and periods of assessment. The Stata-Do files that I have used to create the data used in this paper are publicly available, which will aid those who want to use the data in future.

The IRP5 data can be used to measure inequality in labour earnings for employees, and I have briefly shown how some of the decisions that are required impact on inequality measurement. The IRP5 data include only employees; overall labour income inequality cannot be estimated without the incorporation of the ITR12 assessment data, which include the self-employed. The IRP5 data also include non-labour income, which could be added to enable estimates of overall income inequality. Even with the ITR12 data, however, the SARS individual level data exclude anyone working outside a tax-registered firm, and so cannot be used to estimate overall inequality in (labour) income.

The IRP5 is an exciting source of data that can shed light on a number of aspects of the labour market in South Africa. This paper is an attempt to provide guidance for researchers wanting to use the data for labour market analysis, which is not a simple task.

References

Bassier, I. (2019). 'The wage-setting power of firms, rent-sharing, and monopsony in South Africa'. WIDER Working Paper 2019/34. Helsinki: UNU-WIDER.

Bhorat, H., M. Oosthuizen, K. Lilenstein, and F. Steenkamp (2017). 'Firm-level determinants of earnings in the formal sector of the South African labour market'. WIDER Working Paper 2017/25. Helsinki: UNU-WIDER.

Cassim, A., and D. Casale (2018). 'How large is the wage penalty in the labour broker sector?' WIDER Working Paper 2018/48. Helsinki: UNU-WIDER.

Ebrahim, A., and C. Axelson (2019). 'The creation of an individual panel using administrative tax microdata in South Africa'. WIDER Working Paper 2019/27. Helsinki: UNU-WIDER.

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

Kerr, A. (2018). 'Job flows, worker flows and churning in South Africa'. *South African Journal of Economics*, 86(S1): 144–61.

Kerr, A., and M. Wittenberg (2017). 'Public sector wages and employment in South Africa'. REDI3x3 Working Paper 42.

Kerr, A., and M. Wittenberg (2019). 'Earnings and employment microdata in South Africa'. WIDER Working Paper 2019/47. Helsinki: UNU-WIDER.

Kerr, A., D. Lam, and M. Wittenberg (2019). Post-Apartheid Labour Market Series [dataset]. Version 3.3. Cape Town: DataFirst [producer and distributor].

National Treasury (2016). 'Explanatory memorandum on the Taxation Laws Amendment Bill 17b of 2016.' Available at: www.sars.gov.za/AllDocs/LegalDoclib/ExplMemo/LAPD-LPrep-EM-2016-02%20-%20EM%20on%20the%20Taxation%20Laws%20Amendment%20Bill%2017B%20of%202016%2015%20December%202016.pdf (accessed 28 April 2020).

National Treasury and UNU-WIDER (2019). IRP5 2011-2017 [dataset]. Version 0.6. Pretoria: South African Revenue Service [producer of the original data], 2018. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized dataset].

Pieterse, D., E. Gavin, and F. Kreuser (2018). 'Introduction to the South African Revenue Service and National Treasury firm-level panel'. *South African Journal of Economics*, 86(S1): 6–39.

SARS (2018). 'SARS Business requirements specification: PAYE employer reconciliation' (August 2018 release). South African Revenue Service. Available at: https://www.sars.gov.za/AllDocs/Documents/PAYE%20BRS/SARS_PAYE_BRS%20-%20PAYE%20Employer%20Reconciliation_V17%200%202.pdf (accessed 28 April 2020)

SARS (n.d.). 'Glossary'. South African Revenue Service. Available at: www.sars.gov.za/Pages/Glossary-G.aspx (accessed 28 April 2020)

Stats SA (2019a). 'Quarterly employment statistics details breakdown' (Excel spreadsheet) Available at: www.statssa.gov.za/publications/P0277/QES_Details_BreakDown_200909_201906.xlsx (accessed 15 November 2019).

Stats SA (2019b). 'Quarterly employment statistics Q1:2019'. Available at: www.statssa.gov.za/publications/P0277/QES2019Q1_presentation.pdf (accessed 28 April 2020).

Tax Tim (2014). 'How to calculate totals under source codes 3696, 3697 and 3698'. Website. Available at: https://www.taxtim.com/za/blog/how-to-calculate-totals-under-source-codes-3696-3697-and-3698 (accessed 28 April 2020).

Tax Tim (2018). 'How to calculate totals under source codes 3696 and 3699 for 2017'. Website. Available at: https://www.taxtim.com/za/blog/how-to-calculate-totals-under-source-codes-3696-and-3699-for-2017 (accessed 15 November 2019).

Wittenberg, M. (2017). 'Measurement of earnings: Comparing South African tax and survey data'. REDI3x3 Working Paper 41.

Appendix

Table A1: Labour income source codes from Pieterse et al. (2018) and Kerr (2020)

	Pieterse et. al. (2018)	Kerr (2020)	Description
1	3601	3601	Income – PAYE
2	3605	3605	Annual Payment – PAYE
3	3606	3606	Commission – PAYE
4	3607	3607	Overtime – PAYE
5	3615		Director's remuneration
6	3616		Independent contractors' income
7	3701	3701	Travel allowance – PAYE
8		3702	Reimbursed travel allowance – PAYE
9	3703	3703	Reimbursed travel allowance – Excl
10		3704	Subsistence allowance local travel – IT
11	3707	3707	Share option exercised – PAYE
12		3708	Public office allowance – PAYE
13		3709	Uniform allowance
14		3710	Tool allowance
15		3711	Computer allowance
16		3712	Phone allowance
17		3713	Other allowances – PAYE
18		3714	Other allowances – Excl
19		3715	Subsistence allowance foreign travel – IT
20	3717	3717	Broad-based empoyee share plan – PAYE
21	3718	3718	Vesting of equity instruments – PAYE
22		3751	Travel allowance – foreign service income
23		3752	Reimbursed travel allowance – foreign service income
24		3753	Foreign reimbursive travel allowance
25		3757	Share option exercised – foreign service income
26		3763	Other allowances – foreign service income
27		3764	Other non-taxable allowances – foreign service income
28		3765	BBE share plan – foreign service income
29		3768	Vesting of equity instruments – foreign service income
30	3801	3801	General fringe benefits – PAYE
31	3802	3802	Use of motor vehicle acquisition by employer, not lease – PAYE
32		3803	Use of asset – PAYE
33		3804	Meals, etc. – PAYE
34		3805	Accommodation – PAYE
35	3808	3808	Employee's debt – PAYE
36	3809	3809	Taxable bursaries or scholarships – PAYE
37	3810	3810	Medical aid contributions. – PAYE
38	3813	3813	Medical services costs – PAYE
39	3814	3814	Non-taxable benefit in respect of NSF pension benefits paid by the employer
40	3815	3815	Non-taxable bursaries or scholarships – Excl

	Pieterse et. al. (2018)	Kerr (2020)	Description
41	3816	3816	Use of motor vehicle acquisition by employer by lease – PAYE
42	3820	3820	Taxable bursaries or scholarships (FE) – PAYE
43	3821	3821	Non-taxable bursaries or scholarships (FE) – Excl
44		3852	Use of motor vehicle acquisition by employer, not lease – foreign income
45		3855	Foreign accommodation
46		3856	Foreign free or cheap services
47		3858	Foreign employee's debt
48		3860	Medical aid contributions – foreign service income
49		3863	Medical service costs (PAYE) – foreign service income
50		3865	Non-taxable bursaries or scholarships – foreign services

Source: author's construction.