UNU WIDER

2
25

Guide to Pay As You Earn data in the National Treasury Secure Data Facility of South Africa

Bruce McDougall and Noreen Kajugusi*

June 2025

WIDER TECHNICAL NOTE

wider.unu.edu

United Nations University
World Institute for Development Economics Research

Abstract: This technical note describes how Pay As You Earn (PAYE) data is prepared for research in the National-Treasury Secure Data Facility, and how each new version is checked for consistency. The information is intended to benefit researchers using this data and curators of PAYE data in other data labs around the world.

Key words: Pay As You Earn (PAYE), administrative data, tax data, data preparation, data quality

JEL classification: C55, C81,J3, J31

Acknowledgements: Thanks to Michael Kilumulume, Andrew Nell, Aimable Nsabimana, and Cephas Musafiri for their comments and inputs.

* UNU-WIDER (at the time of writing); corresponding author: **bruce.mcdougall@wider.unu.edu**

This study is published within the UNU-WIDER project **Southern Africa—Towards Inclusive Economic Development (SA-TIED)**.

https://doi.org/10.35188/UNU-WIDER/KGTD3061

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

1 Introduction

In a collaborative effort between the United Nations University – World Institute for Development Economics Research (UNU-WIDER), the South African National Treasury (NT), and the South Africa Revenue Service (SARS), administrative tax data has been shared for research in a secure data facility known as the National Treasury Secure Data Facility (NT-SDF).

One of the datasets available to researchers is the Pay As You Earn (PAYE) data, which is populated by employers on behalf of their employees. In South Africa, the tax form used to capture PAYE information is called the IRP5, and so colloquially this dataset is known as 'the IRP5'. The more formal name we use for the dataset is the Employee Income Payroll Certificate or 'EIPC'.

This note explains how the EIPC data is prepared and how each new version is checked for consistency. We expect this information will benefit two audiences. Firstly, researchers can use it to gain a better understanding of the data, how it is created and how it can be used. Secondly, while we focus on our experiences with South African PAYE data, we believe the lessons generalize to other research labs looking to share PAYE or personal income tax (PIT) data. The current version of the dataset is called e6-v1, as it is the first release of the EIPC based on the sixth annual extraction of raw data received from SARS (National Treasury and UNU-WIDER 2024).

Before proceeding, it is helpful to note that there are several datasets in the lab related to the EIPC, which can make things confusing. The first of these is the raw EIPC data from the tax forms, which is loosely structured and, therefore, not released to researchers. The most basic version of the data that we do release is called the EIPC-ANU (Employee Income Payroll Certificate - Annual) because we transform the raw data to have one row per certificate annually. The preparation of this dataset is the subject of this note. However, the EIPC-ANU is then used as input for various other data products. Table 1 below lists the EIPC-related datasets currently in the lab and explains their relationships:

Table 1: EIPC datasets list

	Formal name	Other names	Description
1.	EIPC-ANU	IRP5	Employee Income Payroll Certificate – Annual. The underlying EIPC data transformed to have one row per certificate per year. The topic of this paper.
2.	EIPC-AFP	IRP5 Firm Panel	Employee Income Payroll Certificate – Annualized Firm Panel. The data from (1) aggregated to the firm level with one row per firm per year.
3.	STMFP	CIT-IRP5 Panel	SARS Treasury Matched Firm Panel. The firm-level data from (2) merged with the corporate income tax (CIT) data.
4.	EIPC-ETI	ETI	Employee Income Payroll Certificate - Employee Tax Incentive (ETI). Additional EIPC data captured for firms that receive the ETI.
5.	STMIP	Individual Panel	SARS Treasury Matched Individual Panel. The raw EIPC data combined with data from the ITR12 tax form.

This note is concerned with the creation of item one on the list, the EIPC-ANU dataset. This, in turn, affects items 2–4.

The layout of the paper is as follows. Section 1 deals with concepts, Section 2 discusses data preparation, and Section 3 details how we check the data for consistency.

2 Concepts

There are three sets of concepts we need to understand before we can work with the data: 1) the source of the data, 2) the goals of data processing, and (3) the raw data.

2.1 The source of the EIPC—jobs, certificates, and submissions

To understand how the EIPC data is prepared, we first need to understand how it came to be. The tax forms that underlie the EIPC were designed to ensure that employers are compliant in terms of their obligation to collect and pay PAYE tax on behalf of their employees. The forms used are the IRP5 and the IT3a: the IRP5 is used if tax is due for the employee in the given tax year, while the IT3a is used if no tax is due (Rossouw 2017). In South Africa the tax year runs from 1 March to 28 February of the following year; the tax year is named after the calendar year in which it ends. Practically, the IT3a and IRP5 forms look the same, and as such, people often refer to either as ‘the IRP5’. As in other countries, PAYE is a withheld tax, meaning a portion of the salary/wage that is kept aside and paid directly to SARS.

In terms of what is captured, the forms are designed to record incomes and financial amounts arising from a job. Various types of remuneration are present (wages, bonuses, fringe benefits, allowances) as well as contributions, deductions, and taxes withheld. Note that these are only recorded if they are related to the job(s) in question. Put differently, the IRP5 is not designed

to capture all forms of income but only those arising from a job. This stands in contrast to the ITR12 tax form, which targets all streams of income (Kerr 2020: 2).

In a given tax year, a person may have multiple tax certificates submitted on their behalf. Each employer they worked for will submit at least one certificate, and in some cases, an employer may issue multiple certificates for the same worker. Instances of multiple certificates may relate to seasonal workers (in farming, tourism, or retail), contract and project workers (such as freelancers, film extras, or event staff), casual workers (like tutors and drivers), interns or other types of short-term work. There is no clear directive from SARS for submitting when employees have more than one bout of work in the year (SARS 2025b). In practice, the decision of whether to submit a single or multiple tax returns for an employee is governed by the employers' practical and operational considerations, which are likely to differ from place to place. There is no way to identify in the data if an employer submitted a single consolidated tax return for multiple short bouts of work from the same worker.

It is useful to draw a distinction between a certificate and a submission of a certificate. A certificate records the relevant information about a job over a certain period. When the certificate is revised, a second submission of the same certificate is created, covering the same period but with updated fields. In this situation, we will see two complete copies/submissions of the certificate in the data, differing only by the updated fields. Because we format the data to be certificate-level (having one copy of each certificate in each year), a key step will involve removing older submissions of the same certificate. As a result of this structure, researchers looking to do person-level analysis will have to further aggregate the data, or turn to a different dataset, such as the SARS-Treasury Matched Individual Panel.

2.2 Why we change the data

The second thing we need is to understand the motivations of data preparation. An important principle when sharing data for research is to leave the original data as unperturbed as possible, as this reduces the possibility of human error, and leaves power in the hands of the researcher to decide how to approach matters. Notwithstanding, there are at least four reasons the data team might want to alter the data. These are: 1) security, 2) reduction of duplicated effort, 3) improved accuracy of results, and 4) computational efficiency.

The first reason is data security. We take steps at SARS and the NT-SDF to ensure that the data we share have minimal risks of disclosing inappropriate information. The second reason we work with the data is to minimize the duplication of effort. This is possible when there is a common task that many people will want to do. For example, the EIPC needs to be converted to the certificate level. The more computationally intense the task, the more time is saved by doing it once for everyone.

The accuracy of results can also be enhanced when we prepare the data in this way. By giving researchers a common starting point, we can rule out cleaning procedures as a cause of variations in their results. Further, in the cleaning process, the data team can work collaboratively with researchers to create the best approach to certain trickier issues, such as defining earnings in the EIPC. Importantly, however, if cleaning is done to improve accuracy, it is critical that it is well documented so that researchers can evaluate the approach taken (the purpose of this note!). The last reason we work with the data is to improve efficiency, which happens at all stages of the data pipeline to reduce computational load.

2.3 Basics of the raw data

The last thing we need is an understanding of the raw data. In this subsection, we explain the identifiers we have in the raw data, and the structure of the tables.

Identifiers

Identifiers are crucial in allowing us to distinguish one unit (certificate/person/firm/etc.) from another. In the EIPC, we have the following identifiers available:

Certificate ID. Possibly the most important identifier in the EIPC is the certificate ID. Each certificate gets a unique ID, and employers will use this ID to identify certificates, for example, if they want to retrieve or revise them. Because, in many cases, each certificate represents a job, the cleaned EIPC-ANU is sometimes called job-level. However, as we have seen, this is not 100% accurate, as short-term bouts of work under the same 'job' sometimes get split onto multiple certificates, or, conversely, employers may combine different 'jobs' that the same employee has had onto a single submission. So, a single job does not necessarily imply a single certificate.

Submission ID and revision number. As mentioned, employers sometimes revise certificates to fix mistakes or update amounts. In this scenario, they will submit a request for correction, and if granted, they will re-submit the certificate. To keep track of this process, SARS uses another variable called *IRP5IT3aID*, which can be thought of as a submission identifier. When a revision is received, the submission ID will change but the certificate ID will remain the same. Because we want to end up with certificate-level data, we will remove superseded submissions, keeping only the latest revision in each case. We do so using a variable called revision number, which increases by one with each revision.

Person identifiers. We have the following person identifiers in the EIPC: the South African national identity number, passport number, and tax reference number. Note that tax reference numbers, in rare cases, refer to entities that are not individuals, such as trusts or associations. The identity number is mandatory for South Africans; if the person is not a citizen, their pass-

port number will be used (Ebrahim and Axelson 2019: 10). Because of this, it can be useful to create a derived person identifier, if you want to include all people in the EIPC at the same time.

PAYE reference number. Firms populate the EIPC forms based on what they see on their payrolls. The office that manages the payroll and submits the forms is identified by SARS using a PAYE reference number (*PAYE_ref_no*) that is mandatory to include on each IRP5. In many cases, the firm will have a single PAYE office and, therefore, a single *PAYE_ref_no*. However, larger firms with many branches may have more than one payroll office, thus ending up with more than one *PAYE_ref_no*. For this reason, the *PAYE_ref_no* is sometimes thought of as a branch identifier. However, this is not entirely accurate, as some of these larger firms may still want to centralize all employees (regardless of location or division) onto a single payroll, even if the workforce is quite diffuse.[1] As such, *PAYE_ref_no* will identify the payroll office the employee is registered at, but this does not always reflect the branch they work at.

Firm identifier. In the raw EIPC data, there is no firm identifier; instead, the data team introduces one. This is done via a table called the conjunction table, which is created at SARS and links PAYE reference numbers to Tax Reference Numbers (TRNs). The firm identifier we bring in is called *taxrefno*, not to be confused with the individual-level tax reference number. It is important to understand that this identifier may not always point to a single 'firm', depending on your definition of 'firm'. Firms may register multiple TRNs due to operating via subsidiaries or registering separate entities to perform different functions. Additionally, TRNs exist not only for corporations paying tax via the ITR14 form, but also for tax-exempt institutions (such as public benefit organizations and state-owned enterprises) that submit IT12IEs,[2] and tax-paying organizations like trusts that submit ITR12Ts. Insofar as these organizations also submit IRP5s, they can also appear as 'firms' with TRNs in the EIPC.

The table below summarizes the identifiers we have discussed:

[1] This is referred to as a 'head-office effect'—when a large firm that is geographically widespread appears disproportionally concentrated in one area, due to employees (or some other measure) being registered at a single administrative centre.

[2] For a full list, see **https://www.sars.gov.za/businesses-and-employers/tax-exempt-institutions/**,

Table 2: Summary of identifiers in the EIPC

	Identifier	Variable name	Description
1.	Certificate ID	certificateno	Certificate identifier.
2.	Submission ID	IRP5IT3aID	Submission identifier. One for each submission of a certificate.
3.	Revision Number	revisionnumber	Counter that increases with each revision of a certificate.
4.	PAYE reference No	payereferenceno	Identifier of the PAYE payroll office this employee is registered at.
5.	National ID	idno	South African national identity number.
6.	Passport No	passportno	Passport number.
7.	Employee TRN	incometaxreferenceno	Tax reference number of this employee.
8.	Firm TRN	taxrefno	The firm level tax reference number associated with the PAYE ref no on this submission.

Structure

The information that SARS captures from the raw submissions is shared with the NT-SDF in three tables. The first is the 'main' table, which contains key submission-level information, such as submission ID, certificate ID, timing of employment, and revision number. The PAYE branch that the worker is located at is also included here, which is how the data can be linked to firms. In addition, certain important line items and subtotals from the 'front page' of the IPR5 are also included here, such as gross income and PAYE withheld.

The second table is the amounts table, which contains disaggregated amounts. These are generally incomes but can also be deductions, contributions, and so on. There are four columns in this table—submission ID, tax year, amount, and the type of the amount. For example, submission 1 might be from 2011 and show R10 000 of income type A and R20 000 of income type B. In South Africa, these income types are categorized according to 'source codes', and a master list of these is maintained by SARS (SARS 2025c). There are around 200 source codes observed in the EIPC, and every non-zero amount is recorded per submission, resulting in around 2 billion rows as of extraction 6. These disaggregated amounts can, in theory, be used to recreate the line items on the front page using the correct combinations, although this can be difficult in practice.

Having the amounts stored separately is optimal. Because the table is large, separating it prevents it from being loaded unnecessarily and slowing down the processing of anything else. Because the amounts table is all numeric, this avoids the generic problem of getting an error due to numeric-vs-string mismatch. More importantly, numeric storage is more efficient, and given that this is the bulk of the EIPC data, this is a big improvement in terms of computation resources and processing time.

The third and final table of the EPIC is the 'person' table, which captures information about the employee, such as their birth date and taxpayer type. Receiving this separately from the main

table is not important from our perspective; the data probably arrives this way as an artifact of how it is stored at SARS. We will link all three tables together using the submission ID, combining the main table with person characteristics, and attaching the relevant amounts. Linking it at the submission level makes sure we are not accidentally pooling different revisions together, which would cause double-counting.

3 Data processing

With concepts out of the way, we can discuss data processing.

3.1 Security measures: de-identification, access restriction, and data checking

Security was the first reason we wanted to alter the data, and the EIPC undergoes several security measures before it is finally shared with researchers.

An important early step is de-identification. At SARS, certain direct identifiers[3] such as names and addresses are removed from the data. Other direct identifiers, such as the South African Identity Number (ID) and the Tax Reference Number (TRN), are encrypted/hashed so that the variable remains in the data but is unintelligible. This hash is done with an algorithm that was developed at SARS and is not shared with the NT-SDF. In rare cases, variables (such as location) are aggregated before being shared with us to reduce the potential for inappropriate disclosure.[4] Note that we refer to this process as de-identification, not anonymization. This is because we cannot guarantee that the data will be completely anonymous through these measures alone, or that someone making a concerted effort will find it impossible to eventually identify a firm/person/entity in the data.

Once these steps are taken at SARS, the data are shared with the NT-SDF in master database file (.mdf) format, which we access using SQL Server Management Studio (SSMS). As an additional security measure, we do not share access to this data with the researcher community; access is limited to the data team.

---

[3] A direct identifier is a variable that, in theory, should allow you to immediately identify a person/firm/unit. For example, a national identity number.

[4] In general, whether to remove a variable as opposed to hashing it will depend on a risk/benefit analysis. Variables that have low rates of error/typos and are designed to track people (or, more generally, units) over time are very useful in a panel context, and these are good candidates for hashing. Applying a strong hash can help circumvent or at least mitigate security concerns while preserving the panel nature of the original variable. However, variables that are prone to error and typos, such as names and addresses, are much more problematic once hashed, as the errors become undetectable. In these cases, it is better to work with them in their original format (in our case at SARS) as necessary.

Before we process the data, we check that the tables received are correct (we have not been given the wrong thing) and that the hashing algorithm has been applied correctly. We then convert it to Stata format and check all values to make sure nothing identifying has been entered erroneously. For certain categorical variables like province or taxpayer type, because we know which values to expect, we check them automatically using fixed lists to make sure no unexpected values have arrived. For the remainder of the text-based variables we tabulate and check them manually. Once we are satisfied, we share the final versions with researchers.

Naturally, these are only the security steps deemed necessary in our context. Other labs might need to be more or less cautious, depending on the sensitivity of their data, the legal context, the attitudes of the data suppliers, and the views of the relevant stakeholders.

3.2 Data preparation

With security protocols in place, we can discuss data preparation.

Step 1: Harmonize names and values

The variable names in the EIPC have changed over the years, sometimes spuriously and sometimes as a result of changes to the underlying IRP5/IT3a form. In cases where the content of the variable is not materially changed, we harmonize the variable names. For example, the variable might have been *amount_sourcecode* in 2011–2015 but was renamed *sourcecode_amount* in 2016 onwards. We do this in the main, amount, and person tables.

Because the data are large, we tend to process them one year at a time, which can make the variable name harmonization tricky. To get it right, it is necessary to load all years, even if you just load a single row per year. In rare cases, we also harmonize variable values. For example, the gender variable is coded as 'F' and 'M' in some years and 'Female' and 'Male' in other years. We convert them all to be 'F' or 'M'.

Step 2: Keep the most recent submission of each certificate

The next step is to keep only the latest revisions. We do this by keeping the most recent submission of each certificate in the main table (based on revision number) and then dropping unmatched submissions from the other tables. Doing this early is preferable as it speeds things up going forward. This results in the dropping of around 5% of submissions in the main table. However, this figure differs quite a lot depending on year.[5] Table 3 below shows a hypothetical example of what this looks like—submission WW of certificate JJ is dropped:

[5] The early years (2008–2011) had an unusually high proportion of submissions that were revised. SARS colleagues have suggested this is related to the fact that the system was brought online in 2008.

Table 3: Main table—removing superseded submissions

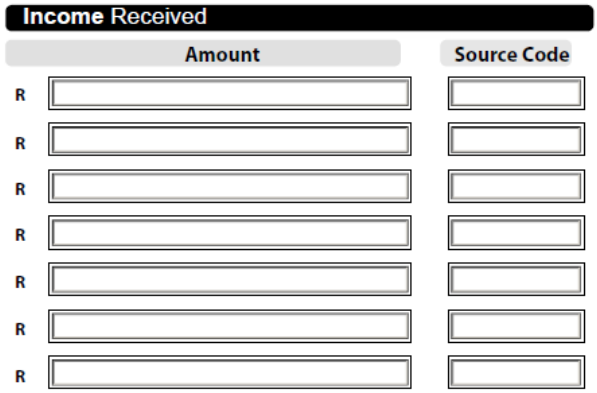
Tax year	Certificate ID	Submission ID	Revision	Employed_from	Employed_to
~~2010~~	~~JJ~~	~~WW~~	~~1~~	~~2009/03/01~~	~~2010/02/28~~
2010	JJ	XX	2	2009/03/01	2010/02/28
2010	QQ	YY	1	2009/03/01	2009/09/30
2010	PP	ZZ	1	2009/10/01	2010/02/28

Step 3: Deal with duplicated source codes in the amounts table

We will match the data in the amounts table to the main table using the submission ID. To simplify this, we process the amounts table to have one row per submission, giving us a neat 1:1 match. This will require first removing duplicated source codes (step 3) and then reshaping to a wide format (step 4).

Duplicated rows arise in the amounts table due to the structure by which source codes and amounts are collected on the forms. For example, the section that captures income-related source codes looks as follows:

Figure 1: How incomes are collected on the EIPC forms



Source: authors' snapshot from a scanned IPR5 certificate.

Each row captures an amount of income and an associated source code. Because each submission refers to a certain employment period, these pairs of amounts and codes reflect the incomes over that period. There is a bit of a problem, however, as the form does not make clear what to do when the person earns the same type of income twice or more within that period. This could be a short-term worker with a consolidated IRP5 submission (as discussed above) or a permanent worker who, for whatever reason, receives the same type of income in batches. Because of this structure, submissions sometimes have multiple rows for the same source code. This problem can arise in the incomes section or the deductions and contributions section, which share the same design.

Table 4 below provides an example of the problem in the EIPC data:

Table 4: Amounts table—submissions with duplicated source codes

Tax year	Submission ID	Amount	Code
2010	XX	50 000	**3601**
2010	XX	50 000	**3601**
2010	XX	7 000	4001
2010	YY	500 000	**3601**
2010	YY	2 000	**3601**
2010	ZZ	3 000	3601

Submissions XX and YY have additional rows, due to the duplication of the 3601 source code (in bold). There are three ways to deal with this. The first approach is to assume that all amounts have been recorded deliberately and correctly and to sum all the amounts for each source code. Another is to assume that any duplicated code was recorded in error and to choose only one of the relevant amounts to keep. Another method is to only assume there was an error if the code *and* amount are duplicated (i.e. the amount does not change, as in submission XX above on rows one and two). In the EPIC-ANU, we take the first option and simply sum it all up per source code per submission.[6]

The result looks like this:

Table 5: Amounts table—duplicated source codes fixed, long format

Tax year	Submission ID	Amount	Source code
2010	XX	100 000	3601
2010	XX	7 000	4001
2010	YY	502 000	3601
2010	ZZ	3 000	3601

Step 4: Reshape the amounts data wide

Table 5 above represents a marked improvement over Table 4, but we don't yet have the desired one row per submission format. The reason is that the source codes and amounts are being stored in a 'long' format, with each distinct code on a new line. The final step is to reshape it wide. There are a few small operations that are needed in the background to get this right, but the end result is straightforward and looks as follows:

Table 6: Amounts table final structure—submission-level, wide format

Tax year	Submission ID	Amount_3601	Amount_4001
2010	XX	100 000	7 000
2010	YY	502 000	0
2010	ZZ	3 000	0

[6] Because we cannot be sure this was the correct approach, we release a separate file with all the duplicate source codes by themselves with a merging variable. This gives the researcher the option to re-introduce them if they choose. Fortunately, this is a rare problem that affects less than one in a thousand certificates, so ultimately, these considerations do not have a large effect.

This format has one row per submission and allows us to neatly merge these amounts with the main table using the submission ID.

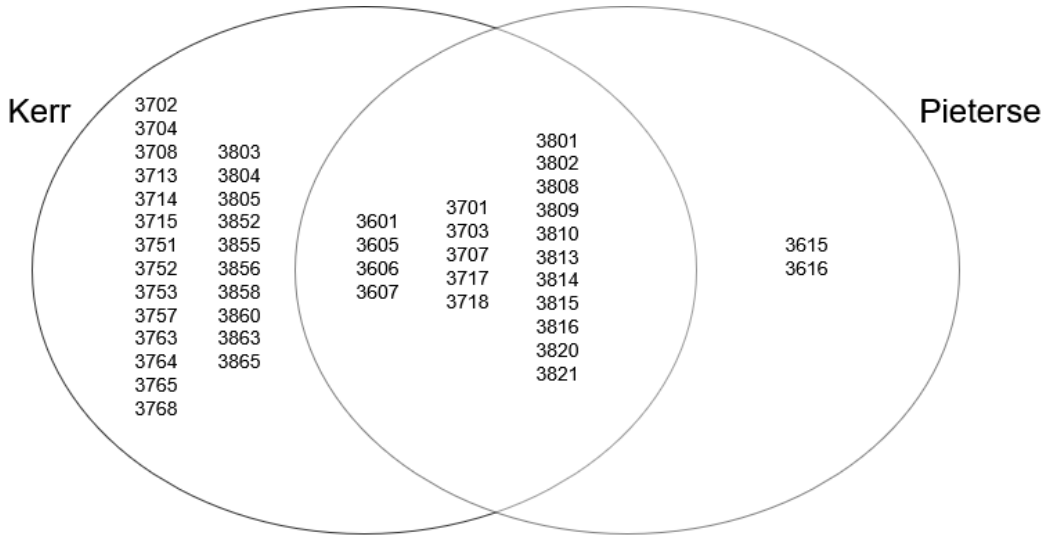
Step 5: Calculate labour income variables

At this point, we are in a position to create a fundamental EIPC-ANU at the certificate level. Before we do, however, let's calculate some derived variables while we are working in the amounts table.

A common need is to have a measure of labour-market income. Researchers want to know the total earnings from labour market employment rather than have it split up over disaggregated source codes. There are various ways to calculate this; for a discussion, see Kerr (2020). The first method is to simply take earnings from source code 3601, which is the 'taxable income' source code. Under this method, all other incomes do not count as labour income. Two alternative methods exist, called the Kerr and Pieterse methods (named after the people who created them). These are different selections of the source codes and whether or not they count as labour-market income.

The figure below summarizes how often the Kerr and Pieterse selections agree or disagree:

Figure 2: Source codes included in labour market income—Kerr vs Pieterse



Source: authors' illustration based on Kerr (2020).

Most readers will be familiar with Venn diagrams. The codes on the left are those that Kerr includes in the income variable, but Pieterse does not. In the middle are the codes they both include, while on the right, we have codes that only Pieterse includes as part of income.

Out of all the source codes that appear in the EIPC, there are 20 that Kerr and Pieterse agree should be part of labour market income, 24 that Kerr counts while Pieterse does not, and 2

where the inverse is true. So, they agree largely, although Kerr's definition is much broader.[7] For a list of these codes and their meaning, see Table A4 in the Appendix. For more discussion of these codes, see Kerr(2020). Basic indicators of whether there was any non-zero labour-market income under each method are also created.

Step 6: Link the tables

The next step is to bring the tables together together. We start with the main table and bring in amounts using the submission ID. Then we bring in the person details in the same fashion, merging on submission ID and dropping the rare (superseded) cases that are unmatched.

Step 7: Bring in additional variables—firm ID and locations

Finally, there are two sets of additional variables we bring in. The first are firm identifiers. As mentioned, because the EIPC contains PAYE reference numbers, we can match these to firm-level tax reference numbers via the conjunction table.[8] Adding firm identifiers is useful for two reasons: it makes it easier if researchers want to merge firm traits into the data, and it makes it simpler down the road when we want to aggregate the EIPC-ANU for use in other datasets, such as the STMFP.

Secondly, we merge in geographic variables. Currently, we receive the geographic data separately from the underlying certificates, so we merge these in on submission ID. We end up with a work or home address for around 85% of the certificates in the data. There are quite a few issues with the quality of these data points however, and so 85% is an overstatement of the amount of usable information in these columns. Note also that we have a separate set of security protocols for researchers who hope to use the geographic element of the data—special permissions are required to analyse these at lower levels of aggregation.[9]

Final structure

Table 7 below shows the final structure of the cleaned EIPC-ANU. We have one row per submission, which is now the same thing as one row per certificate. The amounts have been summed up per source code and appear in wide format, and the derived variables appear on the far right. Because of this format, we describe the data as certificate-level, or occasionally as 'job-

[7] Note that there is a small discrepancy between the codes in Figure 2 and Kerr (2020), in that codes 3709–3012 are excluded in the figure. This was based on looking the historic EIPC-ANU cleaning code, which omitted them (deliberately or accidentally). A priority is to keep the measures consistent, so they remain omitted in this version.

[8] Some processing of the conjunction table is required to treat the rare cases where a PAYE reference number is associated with more than one firm. We start by keeping the most recent tax reference number for each PAYE reference number. In the even rarer cases where there are multiple equally recent TRNs for a PAYE reference number, we keep the modal TRN, choosing at random if there is a tie.

[9] Please see the latest output rules for more, available here: **https://sa-tied.wider.unu.edu/data**.

level' if we are speaking loosely. Note that we omitted the periods of employment (and around 100 other variables) in Table 7 below for readability.

Table 7: Final structure of the EIPC

Tax year	Certificate ID *(certificateno)*	Submission ID *(IRP5IT3aID)*	ID number *(idno)*	Amount 3601	Kerr income	Town
2010	JJ	XX	A	100 000	110 000	Cape Town
2010	QQ	YY	B	502 000	520 000	Mbombela
2010	PP	ZZ	C	3000	3100	Gqeberha

3.3 Outstanding issues

There are a handful of outstanding issues the researcher could tackle to improve their analysis. The first is the question of outliers—we make no effort to flag or remove these in the current version. The second is unrealistic-looking certificates. One example is when there are unfeasibly many certificates for the same person over a given time period, suggesting that they have various overlapping jobs (for the same or different employers), although some of these may legitimately be the types of short-term workers discussed above.

Thirdly, issues have been raised regarding the start and end of employment variables. Some certificates have unlikely or impossible start and end periods of employment. For example, certificates appear with start or end dates that are outside of the tax year in question, which might indicate that employers are submitting the dates of the entire role rather than the reporting period. A less common problem with the dates is cases where the job is recorded as having ended before it started. Using these variables, Kerr (2018) has also pointed out unexpected dips in employment by date of year, further bringing into question how reliable the information is.

Researchers may also want to revise or update the Kerr/Pieterse methods, or create their own method for calculating labour market income. These were created with certain purposes in mind and may not fit all research topics. Further, the source codes change over time, so these concepts will eventually start to 'slide' and become less accurate.

4 Data evaluation

In this section, we will do a common-sense evaluation of e6-v1 to see that it looks sensible and is comparable to previous versions.

4.1 Change happens

From an end-user's perspective, different versions of a dataset that cover the same years should tell the same story. If this is not the case, then there should be plausible explanations, as researchers want to know that their results reflect economic processes rather than spurious variations in data or data cleaning. To this end, we try to maintain consistency from one version to the next. Ideally, we would only adjust the code to add a new year, harmonize variable names, and potentially adjust things mechanically to improve efficiency.

However, change happens, and this sometimes requires changing the code. For example, SARS changed the structure of the EIPC, which was previously stored in one large table and not the three tables described above. Additionally, the amounts previously came in wide and not long format. Over the years, SARS has also added, removed, renamed, and reformatted certain variables. All of these changes make it necessary to update the code so that the output of the data processing remains consistent, even if the input has changed.

Changes to the underlying data points can also happen, and the data team does not attempt to fix these. This happens if taxpayers provide revisions of their submissions or SARS performs its own updates to the data. These changes sometimes take several years to reflect. This is not necessarily a problem but can affect results and aggregates, for example, if a particularly large outlier is added or removed. The datasets also tend to take a while to be fully populated—a dataset that was received in mid 2023 will not be fully populated for the 2023 tax year. Because things change, we perform consistency checks whenever we release a new version to make sure no mistakes were made in data preparation and to confirm that the data looks reasonably consistent from version to version.

In this section, we will compare the present e6-v1 version to data from previous extractions[10] in three ways. Firstly, by examining yearly counts for concepts like the number of submissions and individuals. Secondly, by looking at aggregate measures of key variables, particularly incomes. Finally, we will do a within-submission comparison to check that values on a particular certificate always match values from the same certificate in a different extraction.

Some steps were necessary to preserve comparability between versions. Because older versions of the EIPC-ANU have all revisions (not just the most recent submission of each certificate), we kept all revisions in e6-v1 for this section. For all checks, we restrict our sample to those where the taxpayer type is an individual and drop observations with no normal income (3601). We do not consider 2022 as it was not fully populated in previous versions.

[10] We have recently changed our naming convention. Datasets that were known as version 4 (v4) or version 5 (v5) are now called extraction 4 (e4) or extraction 5 (e5), respectively.

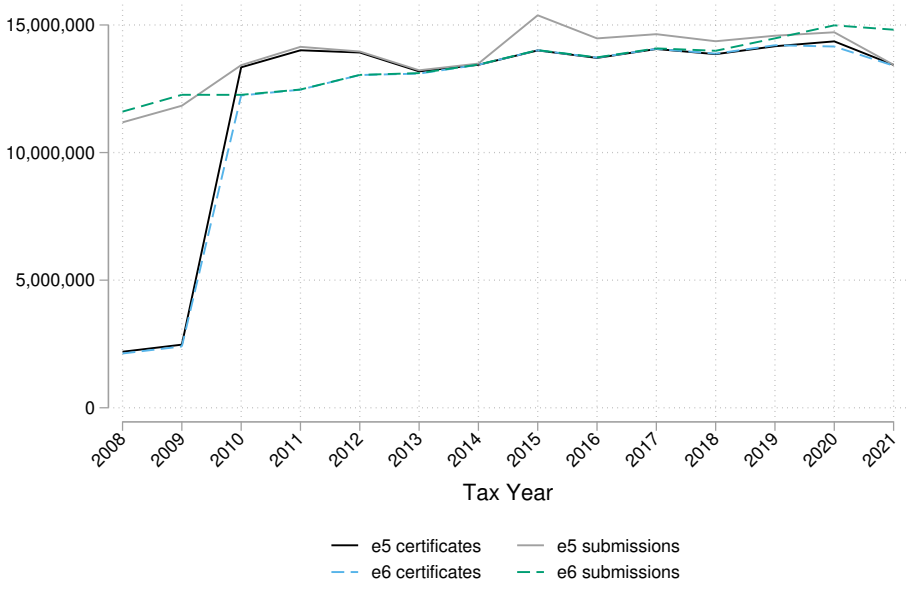
4.2 Comparing counts

Let's begin by comparing e6-v1 to e5 using simple counts.

Certificates and submissions

Figure 3 below shows the number of certificates and submissions in e6 and e5. Certificates are counted by the number of distinct values of certificate ID, while submissions are counted in the same way using submission ID.

Figure 3: Number of certificates and submissions



Source: authors' illustration based on National Treasury and UNU-WIDER (2023, 2024).

The figure tells an encouraging story. The e5 trends closely resemble their e6 counterparts, suggesting strong continuity from one extraction to the next. The number of submissions also tends to track the number of certificates well in both extractions. The overall picture suggests a slow growth in the number of certificates between 2008 and 2021, from 11 million to around 14 million. These counts seem reasonable given what we know about the South African labour market.

Most certificates tend not to be revised, which is why the number of submissions is generally close to the number of certificates. The striking exception is the early years, where the number of certificates is drastically lower, implying a far greater number of revisions per certificate. Colleagues at SARS have suggested that this is an artifact of the period when the data collection system was brought online; exactly how this led to what we see in Figure 3 is unclear. It is not a case of a data error, for example where a single value is repeated millions of times; in the data we see thousands of different certificates with exceptionally many revisions.

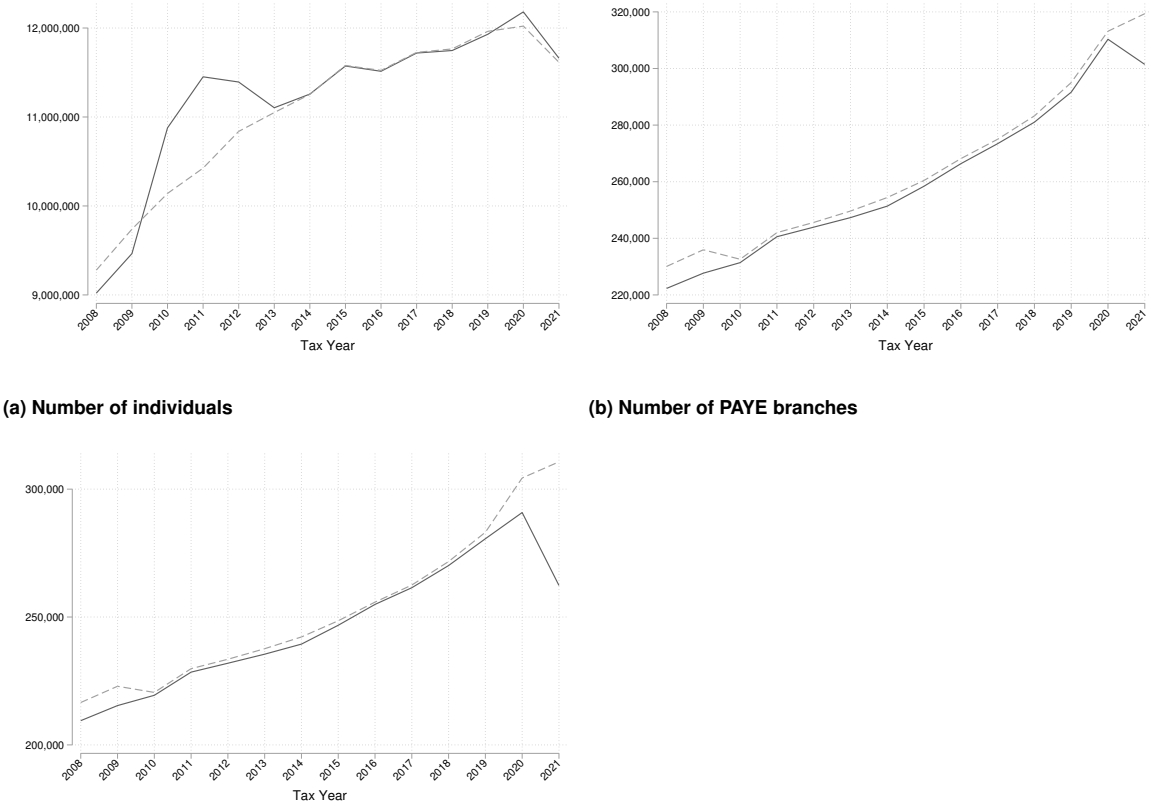
Interestingly, e6 has no revisions between 2010 and 2018. Recall that we did not drop old (superseded) revisions. It seems that SARS may have opted to only keep the most recent revision in these years in e6. Beyond 2018, revisions reappear in e6. By contrast, e5 tends to have revisions across the board, albeit to varying degrees.

Overall, the picture is plausible and tells the story of the digitization of the tax system, followed by a relatively consistent population of submissions and certificates over time. There are minor fluctuations and differences, but altogether e5 and e6 track each other closely.

Employees, PAYE branches, and firms

Next, we count the number of individuals, PAYE branches, and firms. Individuals are counted using a derived person identifier, while branches and firms are measured using their identifiers from Section 2.3. Figure 4 below provides the results.

Figure 4: Counts of key measures



(a) Number of individuals

(b) Number of PAYE branches

(c) Number of firms

Source: authors' illustration based on National Treasury and UNU-WIDER (2023, 2024).

As before, the trends look reasonable and suggest a slow and steady growing of the tax base, up until 2021.

The e6 curves again track the e5 curves closely, with little divergence. There are two exceptions where the curves do diverge. The first is the jump in the number of individuals between 2009 and 2012. It is difficult to know what caused this; the derived person identifier is created similarly in both extractions. The second is the 2021 tax year, where branches and firms drop off in e5. This will partly be due to differences in data maturity: as we saw in Figure 3, a significant number of revisions were added in e6, and these may have been updated to include better information on the branch. The conjunction table was also updated in e6, allowing the data team to better match firm identifiers to the data.

By contrast, in the case of individuals the curves do not diverge in 2021, and both show a sharp decline in the number of employees. It seems that maturity did not affect this trend, and this suggests that something changed in the real world to affect employment. 2021 was an unusual year as the COVID-19 pandemic hit South Africa in March 2020, coinciding with the start of the 2021 tax year. It therefore seems possible that e6 is accurate for 2021, reflecting an unusual pattern: an increase in the number of firms/branches accompanied by a decrease in the number of employees.

Altogether, the results remain encouraging and indicate that e6 is highly comparable to e5. While there are some breaks in trend and a divergence between versions in 2021, we have reasonable explanations for this, and we expect the discrepancy between versions to diminish in future extractions.

4.3 Comparing aggregate income and tax amounts

Having established that e6 closely resembles e5 in terms of counts, the next step is to look at aggregates. We will focus on five key variables: 3601 income, Kerr income, Pieterse income, gross non-taxable income, and PAYE withheld. These we aggregate using totals, ratios, and percentiles. All tables referenced in this section are presented in the appendix.

Totals

Table A1 presents totals for the five variables. Both the e5 and e6 trends exhibit some volatility—for example, total 3601 income starts at 8.5 trillion in 2008, drops sharply to 1 trillion in 2009, and then rises again to around 8 trillion by 2021. One explanation for these fluctuations is the presence of outliers, which have not been removed; a single erroneous entry with extra zeroes could dramatically skew totals. Given this volatility, it is encouraging that e6 still follows e5 closely overall.

The only notable exception is 2021, where totals in e6 are roughly 20% higher than e5, and again this is likely due to differences in maturity. This is quite a substantial change to arise from revisions alone, although, again, outliers may be having a large effect. It is also interest-

ing to note that the totals in e6 are marginally different to e5 in all years. We would not expect employers to revise certificates many years after the fact, so it seems something else may be happening at SARS to either change the values on the submissions, or the population of submissions. We did see above in Figure 3 that e5 tended to have more submissions than e6 before 2020, so it is possible the removal of certain submissions in these years in e6 is causing the changes.

In terms of PAYE and gross non-taxable income, the trends also track each other closely, and these appear more stable year-on-year. The share of PAYE as a proportion of income (calculation not shown) is also stable, generally around 10%, while gross non-taxable income tends to be around 1% of the size of income—these are plausible relationships.

Ratios

Let's compare the relative size of our measures more explicitly. Table A2 shows the ratio of (total) 3601 income to Kerr and Pieterse income. The results are in line with expectations: 3601 income is the most narrow definition of income, so it follows that it is always smaller than Kerr or Pieterse, reflected in a ratio that is less than 1. Generally, 3601 is around 0.8 the size of Kerr and 0.85 the size of Pieterse. Kerr income is larger than Pieterse, which makes sense as it is the broadest of the three.

Once again, 2021 is the unusual year. In this case, the ratios are substantially lower, with 3601 being around 50% the size of the others. Notably, the drop in 2021 is present in both e5 and e6, so this seems to be part of an underlying process and not the result of maturity differences. From Table A1 we know that total 3601 did not decrease significantly, so it must be that one of the other categories that increased.

A policy change may be relevant here. On 1 March 2021, SARS aligned the treatment of provident funds to follow the same 'two thirds' rule that pension and retirement annuities follow. This rule requires that members preserve at least two thirds of their savings in the form of an annuity, with a maximum of on -third available for lump-sum withdrawal (Chong 2021). In other words, members can no longer withdraw the entire amount as a lump sum upon retirement. These annuity payouts are taxed at a higher rate than lump sums and are reported under a new source code (3618) which does not form part of Kerr, Pieterse, or 3601 income.

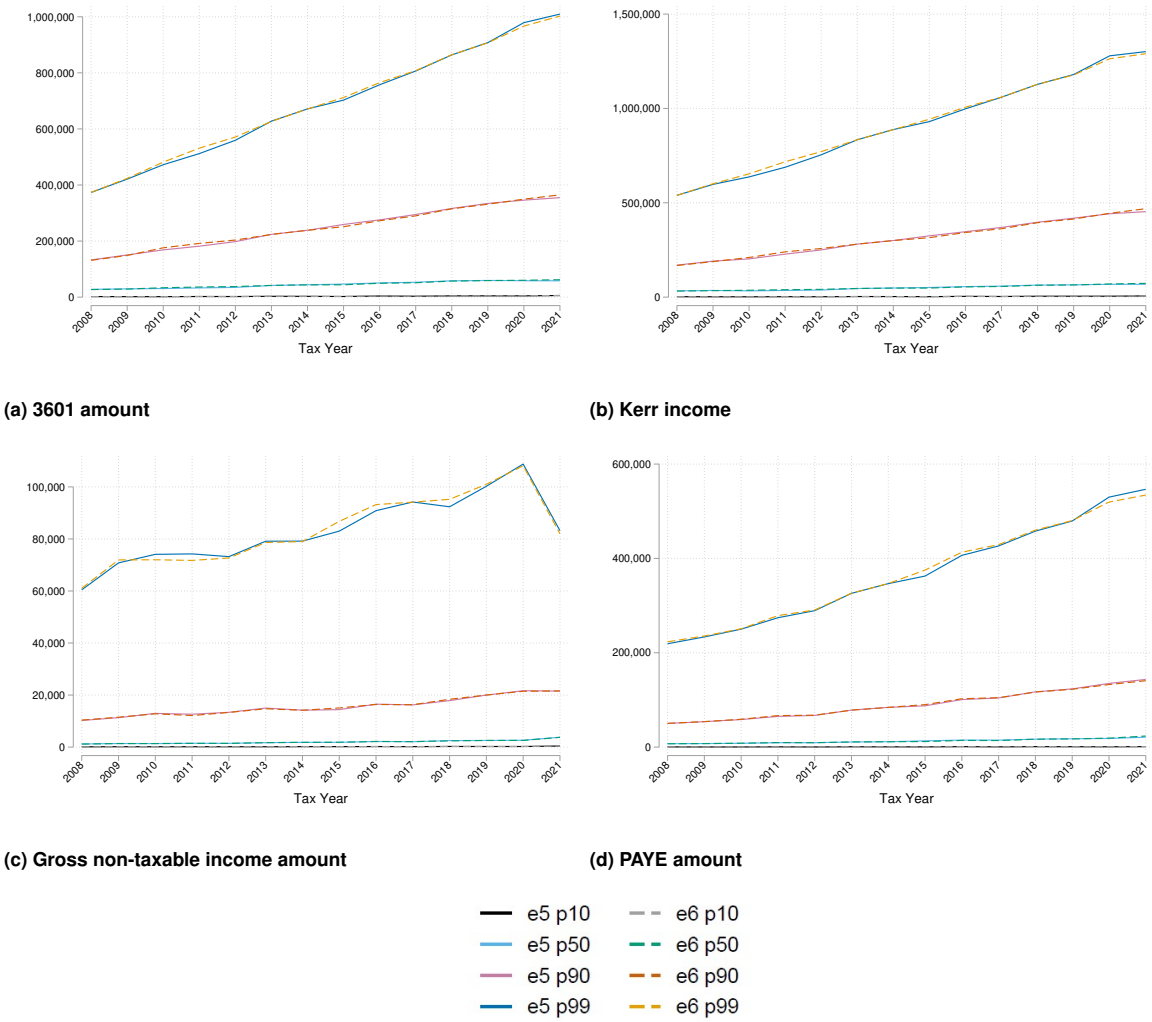
If some of the income that would previously have been taken as a lump sum was directed to other forms of income that form part of Kerr or Pieterse, this might explain why these measures increased relative to 3601 in 2021. Looking in the data (not shown), the 3605 code for annual payments did increase significantly in 2021. It is imaginable that this arose in response to the policy change, with employees (especially those near retirement) asking to be paid in lump sums by their employers under 3605, rather than out of their pension funds, if they wanted

more liquidity immediately. However, this is speculative, and the change could just as easily have arisen from an outlier appearing in e6, or some effect of COVID-19.

Percentiles

Lastly, we examine the variables using percentiles. For the sake of presentation, we skip Pieterse income, as it is highly correlated with Kerr income. We plot the p10, p50, p90, and p99 for the four remaining measures. Figure 5 below shows the results.

Figure 5: Evolution of income percentiles over time



(a) 3601 amount

(b) Kerr income

(c) Gross non-taxable income amount

(d) PAYE amount

Source: authors' illustration based on National Treasury and UNU-WIDER (2023, 2024).

Like our other aggregate measures, the percentiles are remarkably similar between e5 and e6. All of the percentiles seem plausible, growing at a steady rate, and reflecting patterns of income inequality that are well known in South Africa: the very rich (p99) tend to earn more than double the rich (p90), who in turn earn multitudes more than the median (p50) and poorer earners.

A striking feature is the sharp decline of the p99 for gross non-taxable income in 2021, reflected in both extractions. This is interesting and could in theory be a response to the provi-

dent fund policy change, with individuals withdrawing less in 2021. However, such withdrawals have generally been taxable (with certain exceptions), so they should not appear in gross non-taxable income. Still, some amounts previously recorded—correctly or not—as non-taxable may have been shifted to other codes, such as the new annuity code (3618) or as annual payments (3605). It will be interesting to return to this in future and examine why gross non-taxable income seems to have decreased amongst the high earners, and what the relationship is to policy, maturity, and 2021 as an unusual tax year.

Altogether, aggregate incomes in the EIPC-ANU seem consistent between e5 and e6, whether we are using totals, ratios, or percentiles. The final step is to check whether incomes are unchanged at a disaggregated level.

4.4 Comparing within-submission income and tax amounts

Our final check is to assess whether measures are consistent at a within-submission level. We examine whether values change for the same submission across different extractions, which should not occur. We first match submissions across e4, e5, and e6, using IRP5IT3aID as the identifier and dropping any observations that do not appear in all three. We then go variable by variable, calculating the average difference in values between one version and another. This is done pairwise, first comparing e6 to e5, and then e6 to e4. In each case, we exclude rows where the variable is zero or missing in either version.

Formally, for variable X, submission i, and year t, we calculate the average difference in values between extraction 6 and extraction z as:

$$\hat{d}^{\Psi} \frac{1}{N} \sum_{i1}^{N} (X_{i,t}^{e6} - X_{i,t}^{ez}) \tag{1}$$

where N is the number of non-zero, non-missing observations in both extractions, and $z \in \{4, 5\}$ is the number of the extraction being compared to extraction 6.

We then perform paired t-tests to evaluate statistical significance. The test statistic is taken from a two-sided test as follows:

$$t \quad \frac{\hat{d}^{\Psi}}{SE(\hat{d}^{\Psi})} \tag{2}$$

where SE is the standard error of $\hat{d}^{\Psi}$. From this test statistic, we retrieve the associated p-values from the t-table. This method is the same as used in Ebrahim et al. (2021), and we thank Michael Kilumelume for sharing code that we adapted for this purpose.

The results appear in Table A3 below. Each cell of the table contains the test statistics, the number of observations in parentheses and the $SE(\hat{\theta})$ in square brackets, with statistical significance reported in the usual fashion using asterisks. Note that there are many NA values in the table, as e4 and, to a lesser degree, e5 do not cover all years. For e4, we can only compare up until 2018, whereas for e5, we will use up until 2021. For brevity, we only went back to 2015.

To help with interpretation, let's take an example using *kerr_income* for the 2015 tax year, looking at e4 as compared to e6. We have 13,976,708 matched submissions. With a test statistic of 0.015 and a standard deviation of 55, the p-value is large and we can conclude that the change in *kerr_income* between e4 and e6 on matched submissions is not statistically different from zero.

The table shows that values for matched submissions generally do not change between versions, which is what we expect. For gross non-taxable income and PAYE, the differences between e4 and e6 or e5 and e6 were always zero. For the labour market incomes, these differences were very close to zero, and usually statistically insignificant. This implies that the variations we saw in totals must be driven by changes to the population of submissions rather than changes to the values within a submission. This is corroborated by Figure 3, where we saw less total submissions in e6 as opposed to e5. It is not clear why submissions would be added or removed so long after the fact, but fortunately, the overall effect on the aggregates and trends tends to be minimal.

4.5 Summary of data evaluation

The evaluation of the latest EIPC-ANU version was positive: trends in the cleaned dataset were reasonable and largely consistent with earlier versions. The only notable exception was the 2021 tax year, where measures in e6 began to diverge from e5. We argued these differences were most likely due to differences in data maturity, potentially amplified by the effect of the COVID-19 pandemic and a policy change. It will be important to revisit this in future to confirm closer alignment between e6 and subsequent extractions.

5 Conclusion

In this paper, we described how PAYE data is prepared and checked in the NT-SDF. In the preparation section, we focused on the transformation of the data to have one row per certificate per year. This involved removing old revisions and duplicated source codes, and reshaping the data to a wide format. We also detailed the security steps we perform to reduce the probability of disclosing sensitive information.

In the evaluation section, we compared the newest version of the EIPC-ANU (e6-v1) with two previous versions. Overall the trends were highly similar, suggesting a healthy level of continuity from one version to the next. Aggregate trends and within-submission values were consistent over time and across versions. There was some divergence in the latest years and we will pay attention to these going forward. The values also seemed plausible given what we know about the South African labour force. Altogether the results were encouraging and suggest that data preparation is consistent and the underlying data is plausible.

References

Chong, J. (2021). 'Annuitisation of Provident Funds Finally Commences'. Webber Wentzel News, 26 March 2021. Available: **https://www.webberwentzel.com/News/Pages/annuitisation-of-provident-funds-finally-commences.aspx** (accessed March 2025).

Ebrahim, A., and C. Axelson (2019). 'The Creation of an Individual Level Panel Using Administrative Tax Microdata in South Africa'. SA-TIED Working Paper 36. Pretoria: SA-TIED. Available: **https://https://sa-tied.wider.unu.edu/sites/default/files/pdf/SATIED_WP36_Ebrahim_Axelson_March_2019.pdf** (accessed March 2025).

Ebrahim, A., M. Kilumelume, and F. Kreuser (2021). 'The Guide to the CIT-IRP5 Panel Version 4'. SA-TIED Working Paper 203. Pretoria: SA-TIED. Available: **https://sa-tied.wider.unu.edu/article/guide-cit-irp5-panel-version-40** (accessed March 2025).

Kerr, A. (2018). 'Job Flows, Worker Flows and Churning in South Africa'. *South African Journal of Economics*, 86(S1): 141–166. **https://doi.org/10.1111/saje.12168**

Kerr, A. (2020). 'Earnings in the South African Revenue Service IRP5 Data'. SA-TIED Working Paper 116. Pretoria: SA-TIED. Available: **https://sa-tied.wider.unu.edu/sites/default/files/pdf/SA-TIED-WP-116.pdf** (accessed February 2025).

National Treasury and UNU-WIDER (2023). Employee Income Payroll Certificate, Annual. 2008-2022. [dataset]. Version e5-v1. Pretoria: South African Revenue Service [producer of the original data], 2023. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonised dataset], 2023.

National Treasury and UNU-WIDER (2024). Employee Income Payroll Certificate, Annual. 2008-2023. [dataset]. Version e6-v1. Pretoria: South African Revenue Service [producer of the original data], 2024. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonised dataset], 2024.

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. **https://doi.org/10.1111/saje.12156**

Rossouw, C. (2017).'Tax Season: Your Top Tax Investment Questions Answered'. *AllanGray*, 6 October 2017. Available: **https://www.allangray.co.za/latest-insights/personal-investing/tax-season-your-top-tax-investment-questions-answered/** (accessed April 2025).

South African Revenue Service (SARS) (2025a). *Guide for Codes Applicable to Employees' Tax Certificates*. Available: **https://www.sars.gov.za/guide-for-codes-applicable-to-employees-tax-certificates-2025/** (accessed February 2025).

South African Revenue Service (SARS) (2025b). *Guide for Completion and Submission of Employees' Tax Certificates 2025*. Available: **https://www.sars.gov.za/guide-for-completion-and-submission-of-employees-tax-certificates-2025/** (accessed March 2025).

South African Revenue Service (SARS) (2025c). *Find a Source Code*. Available: **https://www.sars.gov.za/types-of-tax/personal-income-tax/filing-season/find-a-source-code/** (accessed March 2025).

Appendix

Table A1: Total of Kerr, Pieterse, and other incomes (in R millions) across e5 and e6 versions

Tax year	kerr income		ptrs income		3601 income		gross non taxable income		PAYE	
	e5	e6	e5	e6	e5	e6	e5	e6	e5	e6
2008	8 520 000	8 535 194	8 493 098	8 507 972	8 321 043	8 331 854	5 348	5 405	117 502	119 765
2009	939 746	962 462	904 445	927 078	717 466	733 668	6 571	6 731	143 253	141 420
2010	1 086 414	1 043 291	1 043 280	1 001 192	890 061	850 268	7 207	7 003	158 666	153 761
2011	1 262 799	1 191 967	1 212 405	1 142 665	1 014 392	947 780	7 161	6 772	186 942	210 844
2012	1 359 526	1 327 489	1 302 655	1 270 982	1 081 550	1 051 480	8 083	8 073	210 187	207 244
2013	1 480 515	1 470 946	1 415 207	1 406 244	1 175 829	1 165 255	8 952	8 850	230 630	228 855
2014	1 905 631	1 894 435	1 834 231	1 823 375	1 573 082	1 563 392	8 942	8 963	255 319	254 514
2015	9 801 427	9 558 345	9 709 602	9 482 119	9 394 177	9 197 734	10 476	9 908	316 690	281 067
2016	2 002 896	1 870 924	1 912 944	1 788 867	1 590 861	1 486 906	11 233	10 664	334 122	313 929
2017	2 199 901	2 064 851	2 104 990	1 976 874	1 741 456	1 633 380	11 434	11 015	363 222	341 648
2018	9 606 525	9 217 858	9 501 639	9 119 892	8 983 939	8 618 712	11 659	11 255	396 056	382 047
2019	2 535 077	2 406 382	2 433 342	2 307 514	2 059 233	1 940 454	12 447	12 298	428 229	419 454
2020	3 169 507	3 248 977	3 064 389	3 142 077	2 592 491	2 660 362	13 559	13 486	476 023	486 541
2021	3 645 700	4 427 424	3 544 270	4 310 501	2 027 115	2 470 942	16 113	17 001	437 659	513 127

Source: authors' compilation based on National Treasury and UNU-WIDER (2023, 2024).

Table A2: Ratio of 3601 income to total Kerr and Pieterse labour incomes across e5 and e6 versions

Tax year	ratio_kerr_e5	ratio_kerr_e6	ratio_ptrs_e5	ratio_ptrs_e6
2008	0.98	0.97	0.98	0.98
2009	0.76	0.75	0.79	0.77
2010	0.82	0.85	0.85	0.89
2011	0.8	0.85	0.84	0.89
2012	0.8	0.81	0.83	0.85
2013	0.79	0.8	0.83	0.84
2014	0.83	0.83	0.86	0.86
2015	0.96	0.98	0.97	0.99
2016	0.79	0.85	0.83	0.89
2017	0.79	0.84	0.83	0.88
2018	0.94	0.97	0.95	0.99
2019	0.81	0.86	0.85	0.89
2020	0.82	0.8	0.85	0.83
2021	0.56	0.46	0.57	0.47

Source: authors' compilation based on National Treasury and UNU-WIDER (2023, 2024).

Table A3: Differences in key variables across e4, e5, and e6 versions

Tax year	Version	kerr_income	ptrs_income	a3601_income	grossntaxableincome	payeamnt
2015	v4	0.015	0.015	0.015	0	0
		(13,976,708)	(13,976,708)	(13,976,708)	(1,446,518)	(7,357,373)
		[55]	[55]	[55]	[0]	[0]
2015	v5	0.015	0.015	0.015	0	0
		(13,976,708)	(13,976,708)	(13,976,708)	(1,446,518)	(7,357,373)
		[55]	[55]	[55]	[0]	[0]
2016	v4	-5.1e-07	-5.1e-07	5.1e-07**	0	0
		(13,597,099)	(13,597,100)	(13,597,096)	(1,375,925)	(7,203,011)
		[0.0043]	[0.0043]	[0.0009]	[0]	[0]
2016	v5	-5.1e-07	-5.1e-07	5.1e-07**	0	0
		(13,698,191)	(13,698,191)	(13,698,186)	(1,381,422)	(7,240,432)
		[0.0043]	[0.0043]	[0.0009]	[0]	[0]
2017	v4	-0.000076	-0.000076	-0.000034	0	0
		(13,878,292)	(13,878,292)	(13,878,284)	(1,407,400)	(7,617,526)
		[0.23]	[0.23]	[0.12]	[0]	[0]
2017	v5	-0.000075	-0.000075	-0.000034	0	0
		(14,042,449)	(14,042,449)	(14,042,436)	(1,423,288)	(7,689,247)
		[0.23]	[0.23]	[0.12]	[0]	[0]
2018	v4	0.01	0.01	0.011	0	0
		(13,427,515)	(13,427,549)	(13,427,450)	(1,299,755)	(7,379,741)
		[40]	[40]	[40]	[0]	[0]
2018	v5	0.0099	0.0099	0.01	0	0
		(13,783,358)	(13,783,392)	(13,783,265)	(1,350,191)	(7,573,823)
		[39]	[39]	[39]	[0]	[0]
2019	v4	NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA

Tax Year	Version	kerr_income	ptrs_income	a3601_income	grossntaxableincome	payeamnt
		NA	NA	NA	NA	NA
2019	v5	-4.1e-06	-4.1e-06	-2.1e-06	0	0
		(14,062,385)	(14,063,330)	(14,060,451)	(1,371,675)	(7,783,892)
		[0.099]	[0.099]	[0.099]	[0]	[0]
2020	v4	NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA
2020	v5	-0.00061	-0.00061	-0.0006	0	0
		(13,806,095)	(13,805,950)	(13,804,519)	(1,342,065)	(7,789,167)
		[1.6]	[1.6]	[1.6]	[0]	[0]
2021	v4	NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA
		NA	NA	NA	NA	NA
2021	v5	0.00061	0.00061	0.000027	0	0
		(12,355,028)	(12,355,027)	(12,353,224)	(1,470,415)	(6,551,678)
		[1.9]	[1.9]	[0.16]	[0]	[0]

Note: the number of observations is reported in parentheses ( ) and the standard errors are in brackets [ ]. ***Significant at 1 per cent, **significant at 5 per cent, *significant at 10 per cent. 'NA' indicates that no comparable field exists between versions of the dataset.

Source: authors' compilation based on National Treasury and UNU-WIDER (2023, 2024).

Table A4: Source codes and meanings

Kerr only:		**Both:**	
3702	Re-imbursive travel allow - tax	3601	Income - taxable
3704	Subsistence allowance (local travel) - taxable	3605	Annual payment - taxable
3708	Public office allowance	3606	Commission
3713	Other allowances - taxable	3607	Overtime
3714	Other allowances - non-taxable	3701	Travelling allowance
3715	Subsistence allowance (foreign travelling) - taxable	3703	Re-imbursive travel allow - non
3751	Foreign travelling allowance	3707	Share option exercised
3752	Foreign re-imbursive travel allow - tax	3717	Employees broad based share plan - taxable
3753	Foreign re-imbursive travel allow - non	3718	Vesting of equity instruments
3763	Foreign other allowances - taxable	3801	Acquisition of assets less than
3764	Foreign other allowances - non-taxable	3802	Use of motor vehicle
3765	Foreign employment subsistence allowance	3808	Payment of employee's debt
3768	Foreign employment vesting of equity instruments - tax	3809	Taxable bursaries and scholarships wrt basic
3803	Right of use of asset	3810	Medical scheme fees fringe benefit
3804	Meal & refreshments vouchers	3813	Medical costs paid by employer iro taxpayer, spouse, children
3805	Free or cheap residential/ holiday	3814	Non-taxable benefit wrt NSF pension benefits paid by employer
3852	Foreign use of motor vehicle	3815	Non-taxable bursaries and scholarships wrt basic education
3855	Foreign free or cheap residential/ holiday	3816	Use of motor vehicle acquired by employer under operating lease
3856	Foreign free or cheap services	3820	Taxable bursaries/scholarships wrt further education
3858	Foreign payment of employee's debt etc.	3821	Non-taxable bursaries/scholarships wrt further education
3863	Medical services costs - foreign	**Pieterse only:**	
3865	Non-taxable bursaries/scholarships wrt basic education	3615	Director's income
		3616	Independent contractors

Source: authors' elaboration based on South African Revenue Service (2025c).