OCTOBER 2020
No. 160

Authors
Rocco Zizzamia
and Vimal Ranchhod
Coordination
Anda David

Research papers

Earnings inequality over the life-course in South Africa

UNION EUROPÉENNE

AFD éditions
AGENCE FRANÇAISE DE DÉVELOPPEMENT

Agence française de développement

Papiers de recherche

Les *Papiers de Recherche* de l'AFD ont pour but de diffuser rapidement les résultats de travaux en cours. Ils s'adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l'ensemble des sujets de travail de l'AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l'ingénieur, sociologie, géographie et anthropologie. Une publication dans les Papiers de Recherche de l'AFD n'en exclut aucune autre.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l'AFD. Ce document est publié sous l'entière responsabilité de son (ses) auteur(s).

Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. AFD Research Papers and other publications are not mutually exclusive.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s).

Earnings inequality over the life-course in South Africa

Rocco Zizzamia
ACEIR, SALDRU - University of Cape Town, University of Oxford
Vimal Ranchhod
ACEIR, SALDRU - University of Cape Town

Abstract
Earnings inequality is usually calculated from a distribution which is measured at a point in time. However, because we typically observe a positive age-earnings profile, a part of cross-sectional inequality is explained by age-related differences in earnings across age cohorts. When inequality is computed using earnings measured over the lifetime, these age-specific differences are averaged out. However, there are also factors that may drive up inequality in earnings measured over time relative to cross-sectional inequality – for instance, low cross-sectional earnings are likely to be correlated with low wage growth and longer spells of unemployment, thereby compounding inequality. Using South African data, we investigate how these dynamic processes act simultaneously but over different time scales to both moderate and exacerbate inequality over time. Because the available panel data in South Africa spans only nine years, straightforwardly constructing a measure of lifetime earnings is not possible. We circumnavigate this challenge by constructing a synthetic lifetime panel by stitching together relevantly similar individuals across successive age cohorts. We use this synthetic panel to compute inequality of lifetime earnings and compare this to inequality of earnings measured over the medium-term (2-9 years), and to inequality measured at a point in time. We find that inequality of lifetime earnings, which reflects the effect of the age/earning relationship, is lower than inequality of contemporaneous earnings. However, inequality of earnings measured over two to nine years, which is more sensitive to inequalities in short-term employment dynamics, is substantially higher than point-in-time estimates.

Keywords
Inequality; earnings mobility; life-cycle dynamics; synthetic panels; South Africa.

Acknowledgements
We are very grateful to Joshua Budlender, Anda David, David Lam and Murray Leibbrandt, and participants at an Agence Française de Développement Inequality Seminar and an African Centre of Excellence for Inequality Research workshop for helpful comments and discussions. The authors gratefully acknowledge funding for this research from the Agence Française de Développement through the EU-AFD Research Facility in Inequalities, a program funded by the European Union. Any errors remain our own.

JEL Classification
J24, J31, J64, O12.

Original version
English

Accepted
September 2020

Résumé

L'inégalité des revenus est généralement calculée à partir d'une distribution qui est mesurée à un moment donné. Toutefois, comme nous observons généralement un profil âge-gains positif, une partie de l'inégalité transversale s'explique par les différences de revenus liées à l'âge entre les cohortes d'âge. Lorsque l'inégalité est calculée à partir des revenus mesurés sur toute la durée de la vie, la moyenne de ces différences spécifiques à l'âge est lisée. Toutefois, il existe également des facteurs susceptibles d'accroître l'inégalité des revenus mesurés dans le temps par rapport à l'inégalité transversale - par exemple, de faibles revenus transversaux sont susceptibles d'être corrélés avec une faible croissance des salaires et des périodes de chômage plus longues, ce qui aggrave l'inégalité. À l'aide de données sud-africaines, nous étudions comment ces processus dynamiques agissent simultanément mais sur des échelles de temps différentes pour à la fois modérer et exacerber les inégalités dans le temps. Les données de panel disponibles en Afrique du Sud ne s'étendant que sur neuf ans, il n'est pas possible de construire directement une mesure des revenus de la vie entière. Nous contournons ce problème en construisant un panel synthétique sur toute la durée de la vie en assemblant des individus similaires de manière pertinente à travers des cohortes d'âge successives. Nous utilisons ce panel synthétique pour calculer l'inégalité des revenus sur la vie entière et la comparer à l'inégalité des revenus mesurée sur le moyen terme (2-9 ans) et à l'inégalité mesurée à un moment donné. Nous constatons que l'inégalité des gains de la vie entière, qui reflète l'effet de la relation âge-gains, est inférieure à l'inégalité des gains contemporains. Toutefois, l'inégalité des revenus mesurée sur deux à neuf ans, qui est plus sensible aux inégalités dans la dynamique de l'emploi à court terme, est sensiblement plus élevée que les estimations à un moment donné.

Mots-clés

Inégalités; mobilité salariale; dynamiques de cycle de vie; panels synthétiques; Afrique du Sud.

Introduction

A well-documented issue in the measurement of inequality relates to the time span over which inequality is being measured. Lifetime earnings inequality is arguably a more relevant welfare concept than inequality of earnings measured at a point in time. In evaluating the disparities in welfare between individuals, we ought to account for both the dis-equalising and equalising forces which play out over time.

This issue becomes pertinent in the labour market, where we typically observe a positive age-earnings profile. Since the workforce is comprised of people of various ages, some component of the inequality in earnings at any particular moment in time simply reflects the difference in the ages of the members of the workforce.

At the same time, however, there are factors that determine inequality other than differences in age at measurement. Inequality of opportunity – determined by location, education, discrimination and many other factors – means not only that different individuals at the same age will have divergent earnings, but also that employment stability and opportunities for earnings growth are distributed unequally. For instance, labour market volatility affects disadvantaged workers more negatively than it does relatively advantaged workers: The least-skilled and lowest-paid workers are more likely to experience more frequent and longer spells of unemployment than their more-skilled and higher-paid counterparts. Even making the unrealistic assumption of stable employment for all workers across the life-course, an additional issue is that earnings growth may differ systematically for those in low wage versus those in high wage occupations.Thus, part of the inequality which is measured cross-sectionally will be compounded over the life-course as the divergence increases over time between workers in low-wage versus high-wage occupations.

These two potentially simultaneous processes exert opposing forces on the difference between lifetime earnings inequality and cross-sectional earnings inequality. On the one hand, when there is mobility in individual earnings over time, cross-sectional inequality includes inequality attributable to age-related differences in earnings. When inequality is computed using earnings measured over the lifetime, these age-specific differences are averaged out. On the other hand, inequality in earnings measured over a longer period captures the effects of inequalities in employment stability or earnings growth. Unlike cross-sectional earnings inequality measures, inequality in lifetime earnings reflects how inequality is compounded over time by these dynamic processes.

The primary objective of this paper is to compare cross-sectional and lifetime measures of inequality in South Africa. In an ideal case, we would have panel data over individuals' lifetimes, in which we observe people as they leave school and enter the labour market, until they permanently exit due to retirement or death. We could then sum earnings over the lifetime and compute inequality based on this lifetime earnings figure. While such long-panel data does not exist in the South African case, the existence of longitudinal data from the nationally representative National Income Dynamics Study (NIDS), which spans a period of approximately nine years, does allow us to estimate a measure of lifetime earnings by constructing a synthetic lifetime panel, subject to a set of stability assumptions.

To do so, we link the experiences of successive cohorts of individuals, such that this forms a "chain" of observations across all age groups, which enables us to generate a proxy for the time-path of earnings across the life-cycle. Using nearest-neighbor matching with replacement, we are able to link the cohort of workers aged between 17 and 22 in 2008 (i.e. who had just entered the labour market in the first wave of NIDS) to a successive cohort aged between 23 and 28 in 2008. This process is then repeated with successively older cohorts until a full synthetic lifetime panel is created. Because with NIDS we have five waves of data for each individual in the balanced panel, we are able to exploit an age overlap between successive cohorts to allow us to match on both time invariant and time-varying characteristics. Thus, by linking relevantly similar individuals across age cohorts we create synthetic individual observations which contain valid responses from approximately ages 18 to 60.

While we are able to construct a synthetic lifetime panel, this by necessity acts as a single birth cohort. This means that, unlike studies using much richer historical data from the US (Bosworth et al., 2001; Haider and Solon, 2006; Kim et al., 2015; Kopczuk et al., 2010; Bowlus and Robin, 2004), we are unable to compare trends in lifetime earnings inequality across age cohorts. In our data, differences between the earnings distributions *between* birth cohorts – which we would expect, for instance, due to educational reform or macroeconomic events – is collapsed through the construction of the synthetic panel. This is a necessary but unfortunate consequence of the synthetic panel strategy that we propose.

The findings of this paper draw on an analysis of the earnings distributions from three versions of NIDS data. First, estimates are provided of earnings inequality using cross-sectional NIDS data. Second, we estimate inequality of earnings summed over multiple periods (between 2 and 9 years) for the same individuals using the genuine 9-year NIDS panel. Finally, we provide the first estimates of inequality of lifetime earnings in South Africa using our purpose-built synthetic lifetime panel.

We find that inequality of lifetime earnings is lower than earnings inequality measured at a cross-section, suggesting that a substantial part of inequality measured at a moment in time reflects age-related inequalities. At the same time, we also find evidence that employment stability is positively correlated with earnings and that employment vulnerability is negatively correlated with earnings. In other words, those who earn low wages when they are employed are more likely to lose their jobs and when they do, tend to experience longer spells of unemployment. Measuring inequality over the shorter age-intervals from the genuine NIDS panel, we find that inequality is higher than measured at the cross-section. Since measuring inequality over these shorter time-periods is less sensitive to aging effects but more sensitive to employment and earnings dynamics, we interpret this as evidence that employment dynamics exert a strong inequality *compounding* effect.

These findings are generally consistent with studies of lifetime earnings and income inequality in the developed world. Bönke et al. (2015) find that, because of earnings mobility over the life-cycle, inequality in lifetime earnings in Germany is about two thirds that computed using annual earnings distributions. Bowlus and Robin (2004) use a shorter-than-lifetime panel and extend this to cover the life-cycle by modelling employment and earnings dynamics for individuals. They find that lifetime inequality in the US

is 40 percent lower than cross-sectional earnings inequality. Kopczuk et al. (2010), using US Social Security Administrative data, find that inequality in long-term earnings, while expectedly lower than inequality in cross-sectional earnings, closely mirrors trends in inequality in cross-sectional earnings. Bowlus and Robin (2012) find that the difference between inequality in lifetime earnings and inequality measured at a point in time is greatest in countries with relatively higher earnings mobility (such as the US), compared to countries with lower earnings mobility (such as much of Western Europe). Flinn (2002) trace the explanation for these differences back to the labour regulatory environment. Where labour regulation is weak (as in the US), employment transitions are frequent, with the consequence that cross-sectional inequality is high but lifetime inequality is relatively equitable. The opposite applies to relatively more regulated labour regimes, such as in Western Europe (Flinn, 2002; Bowlus and Robin, 2004). Tejada (2016) provides a rare and potentially unique set of estimates of lifetime earnings inequality for a developing country: Using Chilean data, he estimates that lifetime earnings inequality in Chile is lower than point-in-time estimates, though remains high.

A key contribution of this paper is to propose a method to construct a synthetic lifetime panel of earnings which incorporates both earnings and employment dynamics, allowing researchers to take both into account in the estimation of lifetime earnings inequality. The method developed in this paper allows for the direct comparison of inequality measured at a point-in-time, over a period of 2-9 years (using the genuine NIDS panel), and over the lifetime (using the synthetic NIDS lifetime panel). This is the first paper to estimate inequality in earnings over the life-cycle in South Africa and one of very few to provide estimates in a developing country context (Tejada, 2016). Our paper thus contributes to a literature which has remained, because of the scarcity of suitable data, dominated by studies focusing on the US and Western Europe.

The paper is structured as follows: The following section elaborates on the motivation for this research and proposes a schematic theoretical framework. The third section discusses data. The fourth section describes the construction of the synthetic lifetime panel. The fifth section presents results. Robustness checks are reported in Section 6 alongside a discussion of several other limitations which are not addressed. The final section offers concluding remarks.

1. Motivation

Inequality has been the subject of much attention in South Africa, which is often considered to be the most unequal country in the world. While the established literature on the topic is expansive, almost all of this work has studied inequality of earnings, income or consumption distributions measured at a moment in time (Leibbrandt et al., 2010, 2018; Tregenna, 2011; Bassier and Woolard, 2018; Leibbrandt et al., 1996; Tregenna and Tsela, 2012; Leibbrandt et al., 2012; Seekings and Nattrass, 2008; Özler, 2007; Sulla and Zikhali, 2018).

While contemporaneous inequality is undoubtedly informative, it is also acknowledged that individual welfare is more meaningfully captured by the expected and realised *evolution* of earnings or income(Sahota, 1978). Because lifetime inequality also depends on the amount and nature of social mobility, annual or cross-sectional inequality is a distorted and limited indicator of lifetime inequality (Friesen and Miller, 1983; Paglin, 1975; Flinn, 2002).

Schumpeter (1955) famously proposed the metaphor of a hotel in which a few people occupy the luxury rooms on the top floors, and many people occupy the more numerous cramped rooms on the lower floors. Inequality measured at a cross-section measures the differences in the quality of rooms on any particular night (Fields, 2006). However, Schumpeter argues that this point-in-time perspective does not account for the potentially equalising role of social mobility - a mechanism through which those guests occupying the top-floor rooms will over time swap rooms with those occupying the rooms in the bottom floors, and vice versa. In the presence of social mobility, cross-sectional inequality is not the only relevant welfare concept in understating social inequities. Lifetime earnings inequality overcomes this issue by incorporating mobility into the measurement of inequality. Using the hotel metaphor, this would be equivalent to measuring the differences between guests in the number of nights spent in rooms of differing quality, over their entire stay.

In the South African literature, some recent attempts have been made to incorporate a dynamic element into the analysis of inequality. Schotte et al. (2018) have, for instance, used differences in the likelihood of falling into poverty to propose a schema of social stratification which is sensitive to the stability or vulnerability of changes in welfare over time. Finn and Ranchhod (2017), using Quarterly Labour Force Survey data, find a modest difference between cross-sectional earnings inequality and a one-year average earnings inequality measure, citing the stability of top earners to be responsible for the small difference. However, given the availability of rich household panel data in South Africa in the form of the NIDS data, there remains scope for more work to be done on inter-temporal inequality.

Before proceeding to our empirical investigation, however, it is instructive to reflect on some stylised scenarios which help to strengthen our grasp of the issues at hand. Consider the following scenarios, represented in the table below:

Case 1 is an economy of two agents (Jack and Jill) spanning two time periods (*t* and *t+1*). In *t*, both Jack and Jill earn $1, and in *t+1*, both earn $2. Earnings in the economy are equally distributed at *t* and *t+1*. In addition, the sum of earnings over both periods are equal for Jack and for Jill.

Case 2 is also an economy of two agents (Jack and Jill), though this time the economy spans three time periods (*t*, *t+1*, and *t+2*). Jack is younger than Jill, and only exists in *t+1* and *t+2*, while Jill exists in *t* and *t+1* but not *t+2*. The economy therefore only consists of two agents at *t+1*, at which point in time Jill earns $2 while Jack earns $1. Measured at the *t+1* cross-section, there is inequality in earnings between the two agents. However, how to interpret the welfare implications of this cross-sectional inequality is not clear since the total earnings of Jack and Jill over their lifetimes is identical.[1]

With these two cases we can see how cross-sectional earnings inequality may be blind to a dynamic issue which has real implications for an inequality assessment. The portion of total inequality attributable to the (arguably more benign) age-dependent inequality affects how we interpret the welfare implications of inequality.

Case 1

	t	*t+1*	Total
Jack	$1	$2	$3
Jill	$1	$2	$3

Case 2

	t	*t+1*	*t+2*	Total
Jack	-	$1	$2	$3
Jill	$1	$2	-	$3

The following two cases illustrate an additional issue - how taking earnings and employment patterns into account when measuring inequality over multiple periods may either compound or mitigate inequality observed at a cross-section, depending on the correlation of earnings with earnings mobility/employment transitions.

Case 3 is an economy with three agents (Jack, Jill and Jim) spanning two time periods (*t* and *t+1*). In *t* Jack earns $10, Jill earns $5 and Jim is unemployed and earns $0. In *t+1* Jack and Jim swap places with Jim now earning $10 and Jack being unemployed. Jill still earns $5. From this example we see that there is inequality in both time periods, but that the pattern of employment dynamics is equalising over time. More precisely, a positive correlation between earnings and the likelihood of becoming unemployed means that high wage earners (Jack and Jim) are likely to experience spells of unemployment, while low wage earners (Jill) are more likely to enjoy greater employment stability. While Case 3 presents this scenario in terms of employment transitions, the same would apply for earnings volatility. Case 3 describes the benign version of social mobility imagined in the Schumpeter hotel metaphor.

Case 4 is also an economy with three agents (Jack, Jill and Jim) spanning two time periods (*t* and *t+1*). In *t* Jack earns $5, Jill earns $10 and Jim is unemployed and earns $0. In *t+1* Jack and Jim swap places with Jim now earning $5 and Jack being unemployed. Jill still earns $10. As in Case 3, we see that there is inequality in both time periods. This time, however, the pattern of employment dynamics is dis-equalising over time because, unlike Case 3, there is now a *negative* correlation between earnings and the likelihood of becoming unemployed.

[1] An additional consideration, not addressed in this paper, is the welfare smoothing effects of credit markets (Kopczuk et al., 2010): if we assume that credit markets are functioning in the economy, it is possible that both Jack and Jill will have been able to smooth *consumption* over the two periods, meaning that (insofar as consumption rather than earnings is the appropriate proxy for welfare) the inequality in earnings at *t+1* will not proxy well for the distribution of welfare.

Low wage earners (Jack and Jim) are likely to experience spells of unemployment, while high wage earners (Jill) are likely to enjoy greater employment stability. Unlike Case 3, we can see that, intertemporally, the negative correlation between earnings and the likelihood of becoming unemployed compounds the inequality observed at either cross-section.

Case 3

	t	*t+1*	Total
Jack	$10	$0	$10
Jill	$5	$5	$10
Jim	$0	$10	$10

Case 4

	t	*t+1*	Total
Jack	$5	$0	$5
Jill	$10	$10	$20
Jim	$0	$5	$5

In practice, however, the processes observed in Cases 1 and 2, and Cases 3 and 4, occur simultaneously. Our prior is that, in most contexts, South Africa included, a positive association between age and earnings in an economy in which individuals' ages differ (as in Case 2) co-occurs with employment and earnings dynamics which are dis-equalising over time (as in Case 4). In the following sections we bring data to bear on this hypothesis.

2. Data

This paper uses Waves 1 through 5 of the National Income Dynamics Study (NIDS). NIDS is South Africa's only nationally representative household panel study, which began in 2008 with a sample of over 28,000 individuals in 7,300 households. There are currently five waves of data available spanning the nine years from 2008 to 2017, where each wave of data is spaced approximately two years apart.

In most of the analysis we use the balanced panel of respondents to exploit the full longitudinal range of the data, restricting our sample to observations which appear in all five waves. In Section 5, however, we pool data from pairs of consecutive waves (t-1 and t), such that the analysis of changes over time represent changes between 2008 to 2010/11, 2010/11 to 2012, 2012 to 2014/15 and 2014/15 to 2017 respectively.

As with any longitudinal study, one is generally concerned with how much attrition there is over time. Starting with Wave 1 in 2008, and including respondents from just the Adult sample, we have a maximum possible sample of 15,597. This number decreases to 13,670 if we exclude adults younger than 25 years or older than 50 years in 2008. Of these, we are able to track only 8,323 in all five waves, giving a net attrition rate of 39.1 percent for this sample. Since it is relevant for the study of inequality, it is worth noting that attrition disproportionately affected high-income earners (Brophy et al., 2018). We use the panel weights released with the NIDS data to correct for the presence of this substantial differential attrition.

3. Creating a synthetic lifetime panel

NIDS is South Africa's only nationally representative household survey panel data set. It spans nine years, from 2008 to 2017. While this makes for richer longitudinal data than is available in most other developing countries, it nevertheless falls short of being usable as a plug-and-play tool for analysing the inequality of lifetime earnings.

However, exploiting the fact that NIDS does have longitudinal data on individuals of different age cohorts spanning nine years, we have developed a method for linking successive cohorts in order to create a single "chain" of individuals. Once linked in the manner that we propose, this "chain" of individuals functions as a synthetic lifetime panel with valid observations from the age of 21 up to the age of 60.

As a first step, we define a set of age variables based on three-year intervals - i.e. ages 20-22, 23-25, etc. The choice of a three year interval was made to maximise the sample size, though a consequence of this choice is that for individuals in the balanced panel – who are observed in five periods – only four age variables are defined. For example, those aged 17 to 19 in 2008 will have variables defined for the following age intervals: 17-19; 20-22; 23-25; 26-28.

The intuition that underlies our strategy for constructing a synthetic panel is to link younger individuals with relevantly similar older individuals, and repeating this process so as to form a "chain" of observations across the life-course. Our method of "linking" younger to older individuals uses nearest-neighbor propensity score matching. This requires us to define variables on which individuals in younger and older cohorts can be matched. To do so, we redefine variables used for matching according to the age-interval variables described above.

Assuming an education variable is one of the variables used in the computation of a propensity score, those aged 17 to 19 in Wave 1 of the NIDS balanced panel will have education variables defined at 17-19; 20-22; 23-25 and 26-28. The same holds true for those aged 20-22 in 2008, except that everything will be shifted across by one age interval: they will have education variables defined at 17-19; 20-22; 23-25 and 26-28. And so on, for those aged 23-25, 26-28, et cetera, in Wave 1.

As one can see in the diagram below, age cohorts defined in this way overlap with successive cohorts by between one and three age intervals. For instance, Cohort 1, aged 17-19, overlaps with Cohort 2 for the age intervals 20-22, 23-25 and 26-28, and with Cohort 3 for the age intervals 23-25 and 26-28. Exploiting this structure, the objective is to match individuals based on shared age intervals, and then append observations from the matched individuals in the successive cohort to those in the original cohort.

		Age						
Cohort	1)	17-19	20-22	23-25	26-28	-	-	-
	2)	-	20-22	23-25	26-28	29-31	-	-
	3)	-	-	23-25	26-28	29-31	32-34	-
	4)	-	-	-	26-28	29-31	32-34	35-37
								et cetera...

Once the initial match is made on these two cohorts, the exercise is repeated for successive cohorts, until we have a continuous "chain" of observations from age of entry into the labour market to age of likely exit. This feature of the data allows us two choices, which involve a trade-off.

First, this overlap of age intervals allows us to match successive cohorts on both time-invariant characteristics (such as gender and race) as well as time-varying characteristics (such as location and employment status). The more periods on which individuals are matched, the greater the probable quality of the match.

Second, just as successive cohorts can be matched on between one and three common time intervals, so can what will become the first iteration of a progressively extended chain of synthetic observations be extended by between one and three non-shared time intervals. The fewer periods on which a match is made, the more periods by which the synthetic panel is extended.

An example will help clarify this: Suppose we match individuals in Cohort 1 to individuals in Cohort 2. By doing so, we are able to exploit the three age intervals for which individuals in these cohorts overlap, and to extend the values of observations for Cohort 1 by one age interval by appending the values of individuals from Cohort 2. Once values assigned to individuals in the original Cohort 1 have been extended by appending values of observations from Cohort 2, the emerging synthetic panel now contains observations for 17-19, 20-22, 23-25, 26-28 *and* 26-28. This exercise is then repeated by matching the existing synthetic panel with Cohort 3, thus extending the dataset by one time period. This process can continue until the synthetic panel has observations from age of entry into the labour market (17-19) until age of exit (approximately age 60).

A choice needs to be made between the following options:

1. *Match-on-1, extend-by-3*: In this approach, we would match on variables shared in *one* period, and build the emerging synthetic panel by extending observations by *three* periods. In the diagram above, this would entail matching Cohort 1 with Cohort 4.
2. *Match-on-2, extend-by-2*: In this approach, we would match on variables shared in *two* periods, and build the emerging synthetic panel by extending observations by *two* periods. In the diagram above, this would entail matching Cohort 1 with Cohort 3.
3. *Match-on-3, extend-by-1*: In this approach, we would match on variables shared in *three* periods, and build the emerging synthetic panel by extending observations by *one* period. In the diagram above, this would entail matching Cohort 1 with Cohort 2.

Each choice involves a trade-off. One consideration is match quality: the greater the number of periods on which two cohorts are matched, the greater the quality of the match on time varying characteristics. However, this comes at the cost of constraining the sample size to fewer initial cohorts. For instance, if, as given in the example above, a choice is made to match on three time intervals, then the "base" of the synthetic panel is restricted to Cohort 1. If, however, a choice is made to match on only one time interval, then Cohort 1 is matched with Cohort 4, thereby freeing Cohort 2 to be matched with Cohort 5, and Cohort 3 to be matched with Cohort 6. By then merging these three separate synthetic cohorts together

(and restricting the lifetime panel to the minimum age of Cohort 3, in this case 23-25), this effectively triples the sample size relative to the "Match-on-3, extend-by-1" strategy.

To balance these two considerations we have elected to use a "Match-on-2, extend-by-2" strategy.[2] This is shown in the stylised figure below, where odd numbered cohorts are matched (in blue), and even numbered cohorts are matched (in green). Nearest-neighbor matching is undertaken on time-invariant variables (gender and race) and time-varying variables (education, location, occupation, employment status, per capita household income and main source of household income). Nearest neighbor matching is undertaken *with replacement*, allowing the algorithm to match a single individual in an older cohort to multiple individuals in a younger cohort. Because of South Africa's youth-heavy age pyramid, younger cohorts in NIDS are substantially larger than older cohorts, making matching with replacement the pragmatic choice.

		Age						
	1)	17-19	20-22	23-25	26-28	29-31	32-34	...
Cohort	3)	-	-	23-25	26-28	29-31	32-34	-
	2)	-	20-22	23-25	26-28	29-31	32-34	35-37
	4)	-	-	-	26-28	29-31	32-34	35-37
								et cetera...

Table 1 reports descriptive statistics for the lifetime synthetic panel (constructed using the "match-on-2, extend-by-2" strategy), and compares these to the NIDS balanced panel. Time-varying characteristics (employment, location, education, household per capita income and main source of household income) are reported in addition to time-invariant characteristics (race and gender). While the balanced panel consists of 8,669 individuals between the ages of 18 and 60, the synthetic panel is built on 1,728 synthetic observations.

The synthetic panel is slightly less urban, less female and more black/African than the balanced panel. It is also substantially poorer - with mean per capita household income being almost half that reported in the balanced panel. The synthetic panel is also on average less reliant on income from the labour market as the primary source of income.

A particular feature of the synthetic panel data is revealed in comparisons in the employment rates and educational attainment in the synthetic panel compared to the balanced panel. While in the balanced panel educational attainment is highest for younger individuals and lower for older individuals (as expected, given the rapid expansion of education amongst the black population in the post-apartheid period), the same pattern is not evident in the synthetic panel, in which the proportion with completed secondary school education *increases* slightly for older individuals. The explanation for this disjuncture can be traced back to the matching algorithm used in the construction of the synthetic panel. Because matches between cohorts are made on a set of variables which include education, secular trends of a population in educational attainment are obscured by the matching procedure which matches highly educated individuals with other highly educated individuals. Since matching is implemented with replacement, highly educated older individuals are matched several times, leading to a "funneling" effect. The same dynamic is evident in employment rates by age – while in the balanced panel employment is highest at age 39 and lowest at ages 27 and 54, in

[2]Sensitivity checks using the other two potential strategies are explored in Section 6.

the synthetic panel, employment rates remain high at age 54. This again reflects the fact that matches between cohorts are made on a set of variables which include employment status.

This feature of the data has implications for how we interpret what *population* is represented by the synthetic panel. The synthetic panel is a composite of multiple age cohorts observed at a cross-section. However, since the matching strategy matches older cohorts to an original cohort (aged approximately 18-21), our strategy in effect simulates the earnings trajectory for this youngest cohort. Since this younger cohort is more educated, it is appropriate that the individuals in the synthetic panel retain this level of education.

Table 1: Sample summary statistics, NIDS balanced panel vs synthetic life-time panels

Balanced panel			**Synthetic lifetime panel**		
	Mean	SD		Mean	SD
Female	0.56	0.496	Female	0.49	0.500
Black	0.85	0.362	Black/African	0.91	0.293
Completed secondary education			Completed secondary education		
Age 27	0.50	0.500	Age 27	0.53	0.343
Age 39	0.42	0.494	Age 39	0.69	0.500
Age 54	0.23	0.419	Age 54	0.67	0.464
Urban	0.60	0.491	Urban	0.52	0.470
Employed			Employed		
Age 27	0.50	0.500	Age 27	0.54	0.498
Age 39	0.63	0.483	Age 39	0.63	0.482
Age 54	0.51	0.500	Age 54	0.67	0.472
Household income per capita (R)	2918	6873	Household income per capita (R)	1735	4700
Main source h'hold income: Labour	0.54	0.498	Main source h'hold income: Labour	0.40	0.490
Observations	8669		Observations	1728	

Notes: The table gives the mean and standard deviation for selected variables from the NIDS balanced panel and NIDS synthetic lifetime panel samples for adults aged 18-60. Summary statistics for the synthetic lifetime panel include observations which are imputed through the matching algorithm which is used in constructing the synthetic panel. These imputations are based on matches (using observable characteristics) between individuals of different ages. Since education is used in this matching strategy, younger individuals are matched with older individuals with similar education levels, which leads to an over-representation of highly educated older individuals in the synthetic panel compared to the balanced panel. "Black" identifies ethnic Africans, which is a dummy variable where other racial categories include White, Indian/Asian and Coloured (mixed-race). "Main source of h'hold income: Labour" is a dummy variable equal to 1 if an individual reports that labour earnings constitute the primary source of household income, and zero if any other income source is listed. Statistics are weighted using the survey design weights, which are corrected for differential panel attrition. The sample is limited to the NIDS balanced panel and to those aged between 18 and 60 years at the time of their interview. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017.

4. Results

4.1. Cross-sectional inequality

In Table 2 we report cross-sectional earnings inequality measured using NIDS data. These results are not restricted to the balanced panel but report cross-sectional inequality for each of the five waves of NIDS. Gini, Theil *T* and Atkinson inequality indices, and 90/10, 90/50 and 50/10 earning percentile ratios are reported. Atkinson and Theil *T* indices and all ratio measures are computed using positive earnings values only, in other words, these measures exclude those who report zero earnings. Gini coefficients are reported both conditioning on positive earnings, as well as where the earnings of the unemployed and economically inactive are coded as zero and these zeros are included in the Gini coefficient. The latter reflects inequality of earnings for the entire working-age population, including those who are not employed at a given time.

For most indicators, we observe a slight drop from the initially high coefficient in 2008, after which inequality remains stable for the following periods. The 90/10, 90/50 and 50/10 ratios indicate the highest levels of inequality for 2008 and 2010/11, after which there is a decline in inequality and a stabilisation. The Gini including zero earnings reflects similar trends. Apart from the high levels of inequality illustrated across all measures, the most salient take-away from Table 2 is the difference between the Gini of positive earnings and the Gini inclusive of zero earnings. While wage inequality is clearly high, when taking into account the contribution that unemployment makes to earnings inequality, the Gini increases dramatically. How these zero-earners are taken into account in the measurement of dynamic inequality will clearly have important implications for the analysis of earnings inequality.

4.2. Employment volatility and inter-temporal inequality

As discussed in Section 2, it is worthwhile in the analysis of inequality over time to investigate whether employment and earnings mobility may compound inequality. We showed how this would be the case if there is a negative correlation between wages and the likelihood of transitioning into unemployment. Table 3 uses the NIDS balanced panel data to show that this is the case in South Africa: Earning low average wages when employed is associated with experiences of long spells out of employment, while those experiencing more stable employment also earn higher wages when employed. The average earnings of those who were employed in all five waves is more than double that of those who were only employed in one of five waves.

Recalling the two stylised cases from Section 2, this shows that South Africa fits the mold of Case 4, in which, all else held constant, employment dynamics exacerbate inequality measured over time relative to inequality measured at a moment in time.

Table 2: Inequality in NIDS cross-sections using various measures

	2008	**2010/11**	**2012**	**2014/15**	**2017**	**Average**
Gini coefficient						
conditional on earnings >0	0.57	0.53	0.52	0.52	0.53	0.53
unconditional on earnings >0	0.84	0.84	0.83	0.81	0.81	0.83
Theil *T* index	0.76	0.55	0.48	0.54	0.41	0.55
Atkinson index (ε=1)	0.49	0.43	0.38	0.40	0.35	0.41
90/10 ratio	13.46	14.66	10.20	12.10	12.00	12.61
90/50 ratio	4.39	4.15	3.52	3.74	4.09	3.95
50/10 ratio	3.06	3.53	2.90	3.23	2.93	3.18

Notes: The table reports Gini, Theil *T* and Atkinson inequality indices, and 90/10, 90/50 and 50/10 earnings ratios for all five waves of NIDS data. All valid adult observations (ages 18-60) for each wave are used rather than the balanced panel. The Theil *T* and Atkinson indices and all ratio measures are computed using only positive earnings - i.e. are limited to those respondents in active employment and reporting a positive wage. Gini coefficients are reported both conditional on reporting a positive wage - i.e. limited to those in employment - and unconditional on reporting a positive wage - i.e. inclusive of those not in employment and reporting zero earnings. The Atkinson index is computed setting the inequality aversion parameter (ε) to 1. The 90/10 ratio is the ratio of earnings of workers at the 90th percentile of the earnings distribution to the earnings of workers at the 10th percentile of the earnings distribution. The 90/50 ratio is the equivalent concept measuring "upper-tail" inequality, while the 50/10 ratio is the equivalent concept measuring "lower-tail" inequality. The final column reports the average of each inequality measure over all five waves. It should be noted that NIDS is arguably poorly suited for measuring cross-sectional inequality: NIDS was launched in 2008 with a nationally representative sample of the South African population at the time. Later waves used the same sampling frame, with the result that the sample remained representative of South Africa's population *in 2008*. In addition, differential panel attrition and the under-sampling of top earners may also have compromised NIDS representativeness over time. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017. Statistics are weighted using the survey design weights.

Table 3: Association between monthly earnings and employment duration in the NIDS balanced panel

Periods employed	Mean earnings (R)	SD	*n*
1/5	3 175	3 533	533
2/5	3 352	3 040	791
3/5	4 835	6 145	874
4/5	6 499	7 194	855
5/5	9 357	9 785	937
all observations	6 004	7 392	3 990

Notes: Data is limited to adult observations (ages 18-60) present in the NIDS balanced panel (Waves 1-5) who report being employed in at least 1 out of the 5 periods in which they were surveyed. The table compares mean monthly wages *during a period of employment* for workers who reported being employed in 1, 2, 3, 4, or 5 periods, out of a possible total of 5 periods. For instance, the first row shows that mean monthly earnings were R3,175 *during the period of employment* for individuals who reported only being employed in 1/5 NIDS waves. Likewise, the penultimate row shows that the mean monthly earnings over all periods of employment for those who were employed in 5/5 Waves was R9,357. The last row reports the mean monthly earnings for all those reporting at least one period of employment in the full balanced panel. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017. Statistics are weighted using the survey design weights, which are corrected for differential panel attrition.

4.3. The age-earnings relationship

Employment dynamics may exacerbate inter-temporal inequality. However, in Section 2 we showed how this may occur simultaneously with the inequality reducing effects of a positive age-earnings profile. If earnings trend upward over the life-course and inequality at a given point in time is measured across individuals of different ages, then when measuring inequality of *lifetime* earnings (thereby netting out age-specific differences), inter-personal inequality will be observed to be smaller than that measured at a point in time, all else held constant.

Figure 1 plots the age-earnings profile for South Africa using NIDS, differentiating between those without a secondary qualification, those with a secondary school qualification, and those with a tertiary qualification. These results are estimated by pooling all of the NIDS cross-sections.

Panel (a) depicts the relationship between age and earnings for those with positive earnings values, i.e. restricted to those in employment. Panel (b) represents the relationship between age and earnings for those with positive *or* zero earnings values. In other words, while Panel (a) plots earnings growth conditioning on positive earnings, Panel (b) additionally accounts for changes in employment rates across the life-course. Note the differences in the vertical axes of the two Panels.

As expected, we observe a positive age-earnings profile for positive earnings throughout the life-cycle. The absolute level of earnings is by definition lower when including the non-employedm as evidenced by difference in the scales of the vertical axes. Until age 40 both lines follow the same trend, at which point the two lines diverge, with a downward slope for earnings inclusive of non-employment. This suggests that from age 50 onward there is a decline in the employment rate rather than in earnings. This graph emphasises the importance of taking both life-cycle dynamics and employment transitions into account when measuring inter-temporal inequality. Figure 1 shows that life-cycle dynamics will introduce age-related inequality into cross-sectional estimates.

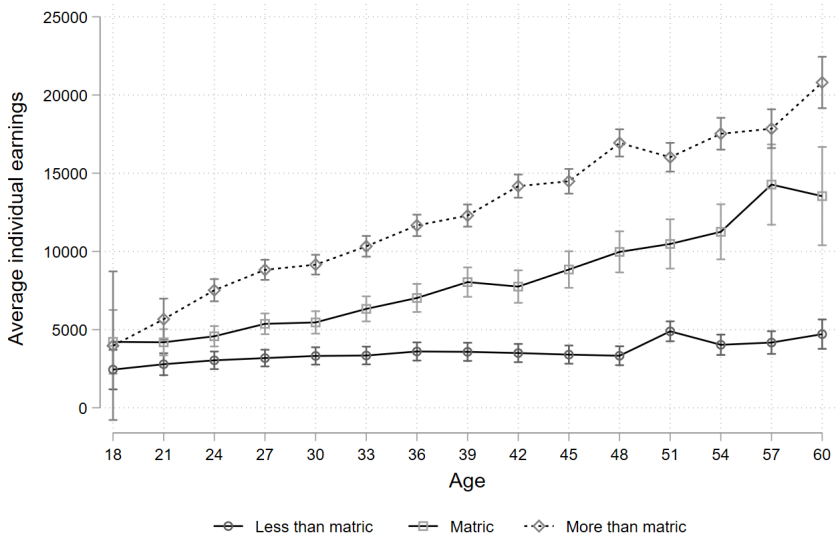
Figure 1 further distinguishes the age/earnings profile by educational attainment. Tertiary- and secondary-educated individuals display a stronger growth trend in positive earnings than those without completed secondary education, for whom there is barely any earnings growth across the life-course. While Panel (b), which takes employment dynamics into account, shows that *employment* rates decline fairly dramatically for tertiary educated (and to a lesser extent secondary-educated) individuals after age 50, across all education categories there is still an aggregate positive age-earnings gradient over the life-course.

4.4. Inequality of lifetime earnings using the synthetic panel

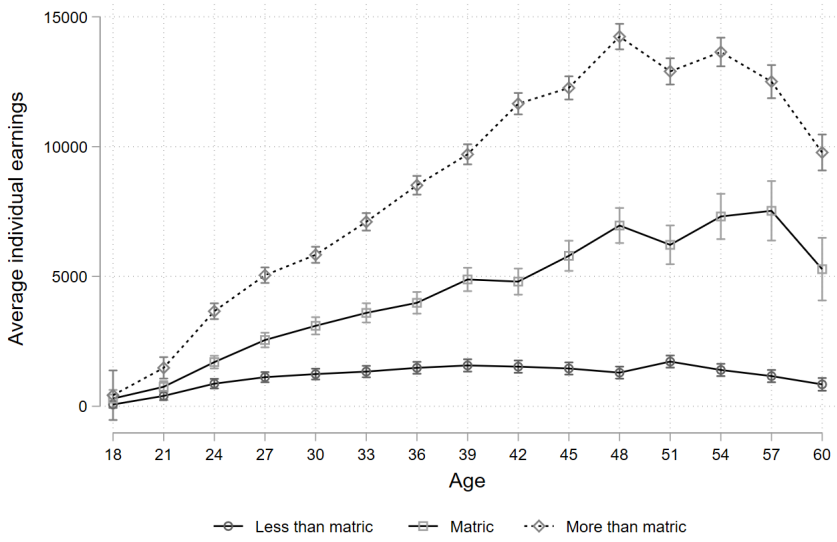
The main results of this paper are presented in Table 4. In this table we compare inequality measures computed using NIDS cross-sectional data, to inequality of earnings summed over several waves in the genuine NIDS panel, to inequality of lifetime earnings in the synthetic NIDS panel (the construction of which is described in Section 4).

Figure 1: The evolution of earnings over the life-course in NIDS cross-sectional data, by educational attainment

(a) Conditioning on positive earnings



(b) *Not* conditioning on positive earnings



Notes: Data is pooled from all five cross-sectional waves of NIDS. Both Panels (a) and (b) report mean monthly earnings across 3-year age intervals, from 18 years to 60 years. Panel (a) computes mean earnings at each age conditional on positive earnings - i.e. is limited to those in employment. Panel (b) computes mean earnings at each age *without* conditioning on reporting a positive wage - i.e. is inclusive of those not in employment and reporting zero earnings at any given age. Panel (a) therefore reports the evolution of *positive earnings* over the life-course, while Panel (b) reports the *combined evolution of earnings and employment dynamics* over the life-course. Statistics are weighted using the survey design weights. The sample is limited to those aged between 18 and 60 years at the time of their interview. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017.

Inequality measures are the same as those reported in Table 2: Table 4 reports estimates of the Theil *T* and Atkinson indices, the Gini coefficient reported both conditional and unconditional on positive earnings[3], and 90/10, 90/50 and 50/10 ratios.

In the leftmost column, inequality estimates using cross-sectional data are reported. These are the average values across the five waves (2008-2017). As noted before, the large difference between the Gini conditional on positive earnings and the Gini unconditional on positive earnings illustrates clearly the centrality of unemployment as a driver of inequality in the labour market.

The middle supercolumn contains inequality measures for medium-term earnings, summed over between two and nine years. For most inequality measures, there is a sharp increase in inequality for medium-term earnings compared to cross-sectional earnings, and this relatively higher inequality continues to rise over two to nine years. For example, the Theil *T* index increases from 0.55 when measured at a point in time, to between 0.60 and 0.72 when measured between 2 and 9 years.

With the exception of the unconditional Gini, all of the other inequality measures do not include the effect of unemployment on *cross-sectional* labour market inequality. However, unemployment *does* enter into the inter-temporal earnings inequality measures, since zeros from periods of non-employment are included when earnings are summed over time. Thus, some of the higher inequality in the inter-temporal measures in supercolumn 2 relative to the cross-sectional estimates is driven by lower-wage workers experiencing spells of unemployment over time.

We might expect that fluctuations in earnings and employment transitions are likely to be important determinants of labour market inequality in the short- and medium-term, while aging dynamics may become more relevant over a longer time horizon. While the former is inequality increasing in South Africa (Table 3), the latter is inequality decreasing (Figure 1).

Measuring inequality over a longer time horizon, however, we might expect that age-related inequalities are averaged out. The final column in Table 4 supports this view: across most inequality measures, inequality measured over the lifetime using the synthetic NIDS panel are lower than both the panel and the cross-sectional estimates. Notable exceptions are the 50/10 ratio and the Atkinson index, where the inequality in lifetime earnings is estimated to be higher than the inequality in cross-sectional earnings.

Our ability to compare earnings inequality over the life-cycle to earnings inequality over the medium-term provides some evidence of how dynamic processes act simultaneously but over different time frames to moderate and exacerbate inequality over time. Because medium-term earnings are subject to employment and earnings volatility but are much

[3]It is important to note that, for Gini coefficients unconditional on positive earnings, when inter-temporal means are computed, zeros at any moment in time will be included in calculating an inter-temporal average. An example is helpful: When computing the 2 Wave Gini (column 2), if an individual is employed in Wave 1 and unemployed in Wave 2, the average is simply earnings in Wave 1 divided by two. This inter-temporal average will be included in the calculation of the Gini conditional on positive earnings (Row 1) and in the calculation of the Gini unconditional on positive earnings (Row 2). However, if an individual has zero earnings in both Wave 1 and Wave 2, their inter-temporal average is zero, and so this inter-temporal average will *not* be included in the calculation of the Gini conditional on positive earnings (Row 1) but will be included in the calculation of the Gini unconditional on positive earnings (Row 2).

Table 4: Estimates of lifetime inequality in South Africa: Comparing cross-sectional, panel and the synthetic lifetime inequality measures

	Cross-section	**NIDS Panel**				**Synth. panel**
	2008-2017	2 Waves	3 Waves	4 Waves	5 Waves	Lifetime earnings
Gini coefficient						
conditional on earnings >0	0.53	0.57	0.59	0.61	0.62	0.42
unconditional on earnings >0	0.80	0.80	0.78	0.77	0.76	0.42
Theil *T* index	0.55	0.60	0.64	0.69	0.72	0.29
Atkinson index (ε=1)	0.41	0.46	0.50	0.53	0.55	0.44
90/10 ratio	12.03	17.68	22.99	26.69	30.28	11.62
90/50 ratio	3.99	4.52	5.07	5.24	5.56	2.38
50/10 ratio	3.01	3.93	4.54	5.09	5.45	4.89

Notes: The table reports Gini, Theil *T* and Atkinson inequality indices, and 90/10, 90/50 and 50/10 earnings ratios for *a)* the NIDS cross-section, in which point-in-time inequality measures are averaged over all five waves, *b)* individual earnings summed over multiple waves (between 2 and 5) in the balanced panel, and *c)* individual earnings summed over approximately 36 working-age years, spanning ages 20-56,21-57 or 22-58 in the NIDS life-time synthetic panel. The inequality measures are as described in Table 2. Cross sectional statistics (supercolumn 1) are weighted using the survey design weights, while for statistics which exploit NIDS's panel dimension (supercolumns 2 and 3) these design weights are adjusted to account for differential panel attrition. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017.

less responsive to aging dynamics, changes in medium-term inequality allow us to partially isolate the contribution of employment and earnings dynamics to inter-temporal inequality. Thus, while lifetime earnings are substantially more equal than cross-sectional earnings, the finding that inequality in earnings over the medium term are more unequal than cross-sectional inequality suggests that labour market dynamics remain a substantial source of earnings inequality.

5. Robustness

There are several limitations and concerns with the use of a synthetic panel in the estimation of lifetime earnings inequality, some of which are unavoidable. Most of these stem from potential shortcomings in the matching strategy underlying the construction of the synthetic panel.

A poorly implemented matching strategy has at least three implications for the measurement of earnings inequality. First, it may be that the earnings trajectories of the constructed synthetic individuals differ *systematically* from the (unobserved) earnings trajectories of the actual individuals whose data is used to construct the synthetic panel. Second, it is possible that while the matching strategy does not introduce bias (as in the first concern), the generated earnings represent little more than random noise. This would lead to a spurious reduction in inequality when measured over time, since random noise would artificially introduce earnings mobility between individuals. Third, it is possible that, when matching *with replacement*, linking individuals across cohorts will mechanically reduce inter-personal inequality in the synthetic panel, as the number of possible earnings trajectories are "funneled" through a reduced number of older panel members.

5.1. Evaluating the performance of the matching strategy in the construction of a synthetic lifetime panel

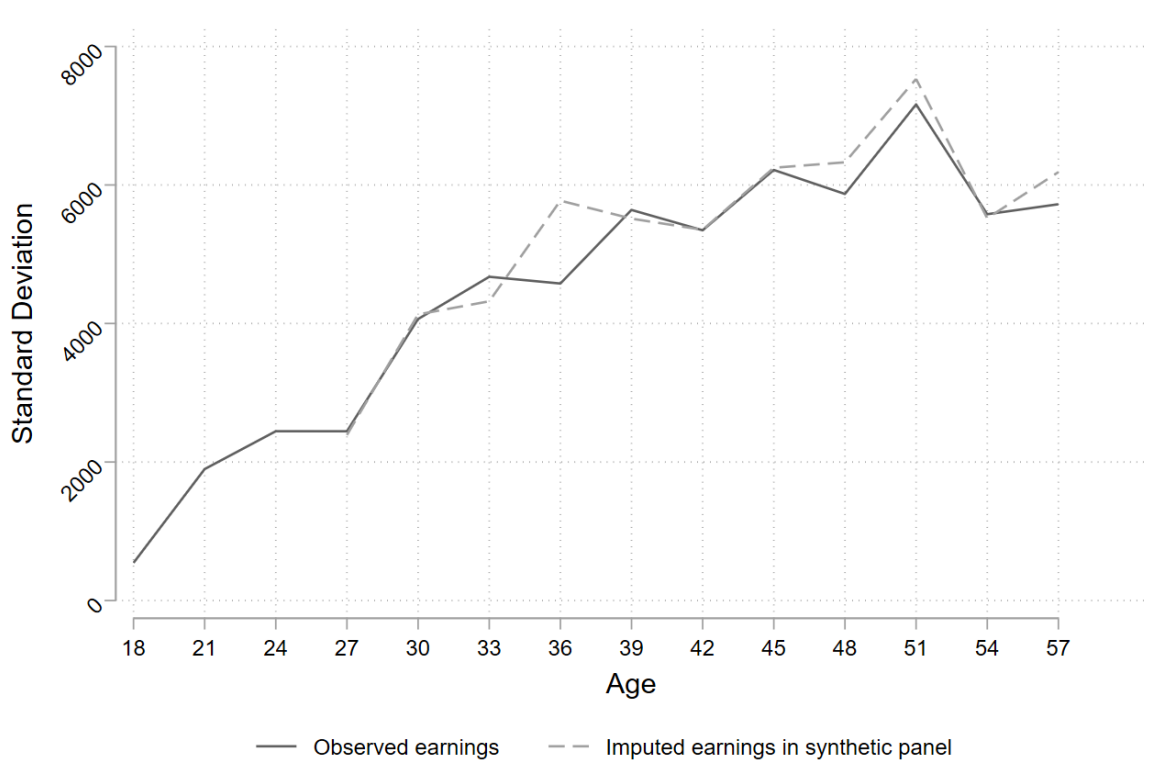
In Figure 2 and Figure 3 we attempt to evaluate the quality of the earnings variables generated through the matching strategy used in the construction of the synthetic panel. Ideally, we would have both observed and imputed (through the match-and-extend procedure) earnings variables for the same individual. This is not the case in the matching procedure as described in Section 4, since earnings are only imputed for periods where earnings were not observed in the original data.

For this reason, we adapt the match-and-extend procedure to allow us to directly compare observed and imputed earnings for the same individual. To do so, we abandon the "match-on-two, extend-by-two" strategy in favour of a "match-on-two, extend-by-*one*" strategy. In this adapted strategy, the periods on which observations are matched is shifted back one period. For instance, individuals who are observed at ages 18, 21, 24 and 27 are matched with individuals observed at ages 21, 24, 27 and 30 on variables which are shared at ages 21 and 24. Earnings at age 27 can then be imputed for the younger cohort from the older cohort (i.e. using a "match-on-2, extend-by-1" strategy). However, observations from the the younger cohort now have both an *observed* wage, and an *imputed* wage (from the matched observations from older cohort), allowing for a direct comparison between the two.

Figure 2 depicts the standard deviation of earnings for observed earnings (solid line) and imputed earnings (dashed line).[4] Standard deviations of both variables track each other closely, illustrating no greater dispersion in the imputed earnings relative to the observed earnings.

[4]Because of the match-and-extend strategy, imputed earnings can only be reported from age 27 onward.

Figure 2: Standard deviation of earnings across the life-course, comparing observed earnings with earnings imputed in the construction of the synthetic panel



Notes: This figure compares the standard deviations of observed earnings (solid line) with the standard deviation of imputed earnings (dashed-line) for the same individuals. To do so, a variant of the matching strategy underlying the construction of the synthetic panel is used. In this case, instead of using a "match-on-2, extend-by-2" strategy, the periods on which observations are matched is shifted back one period, and a "match-on-2, extend-by-1" strategy is used. For instance, individuals which are observed at ages 18, 21, 24 and 27 are matched with individuals observed at ages 21, 24, 27 and 30 on variables which are shared at ages 21 and 24. Earnings at age 27 can then be imputed for the younger cohort from the older cohort (i.e. using a "match-on-2, extend-by-1" strategy). However, the younger cohort now has both an *observed* wage, and an *imputed* wage (from the older observation), allowing for a direct comparison between the two. In the figure, it is only possible to report the standard deviation of imputed earnings for age 27 onward, while the standard deviation of observed earnings is reported from age 18 onward. The sample is limited to the NIDS balanced panel and to those aged between 18 and 60 years at the time of their interview. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017.

Figure 3 depicts the standard deviation of the *differences* between observed earnings compared to imputed earnings for the same individuals. The absolute difference between observed and imputed earnings is computed for each individual where both observed and imputed earnings are available. The straightforward intuition is that the greater the standard deviation in the differences between observed and imputed earnings, the less confidence one would have in the matching strategy.[5] However, beyond this intuition, it is unclear how one ought to evaluate the magnitudes of the statistics reported in Figure 3. The standard deviation of the differences of the observed earnings of the same individual spaced 3 years apart in the NIDS panel (i.e. actual earnings at age 27 minus earnings at age 24, for the same individual) is of a comparable magnitude as those reported in Figure 3, for instance.

What is clear, however, is that the match-and-extend strategy is better at predicting earnings for younger compared to older workers. This is, however, more a function of the greater standard deviation in earnings at older ages (Figure 2) than it is of the robustness of the matching strategy.

5.2. Does the matching process mechanically bias our estimates of lifetime earnings inequality downward?

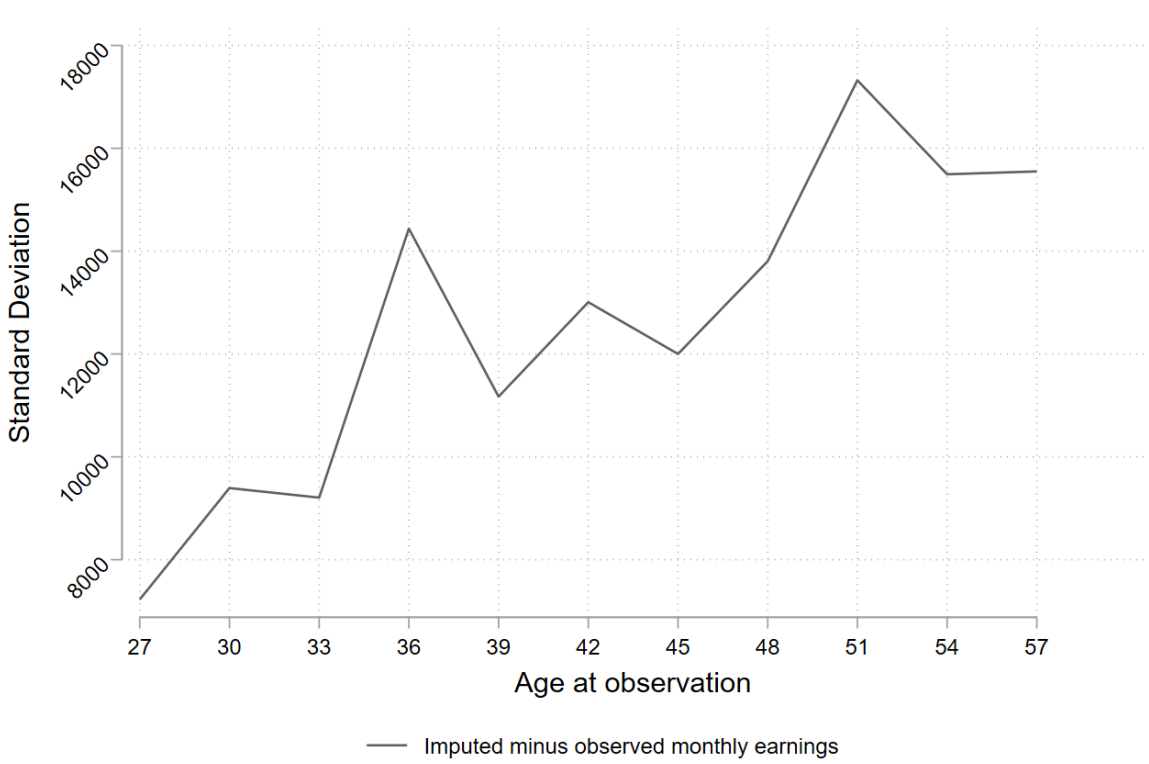
As discussed above, a further concern is that, because matching is undertaken with replacement, that there may be a mechanical reduction in inequality in the synthetic panel relative to a genuine panel spanning the same period. Clearly, testing this directly is not possible, since the synthetic panel data is only necessary because of the absence of a genuine panel.

However, using the same adapted "match-on-2, extend-by-1" strategy described in the preceding subsection, we can test for the presence of this downward bias. To do so, we compute two sets of Gini coefficients of earnings summed over a period of nine years. The first is computed using *only genuine panel data* - i.e. where each of the four discrete points at which earnings are observed are unchanged from the original NIDS data. The second set of Gini coefficients is computed using *partly synthetic panel data* - where three of the four earnings observations are as directly reported by the individual respondent in the NIDS panel, while the fourth earnings observation is imputed using the "match-on-2, extend-by-1" strategy described and used in the preceeding subsection.

In Figure 4, the solid line reports the genuine 9-year Gini coefficient while the dashed line reports the partly synthetic 9-year Gini coefficient. The two lines track each other closely, suggesting that there is no substantial downward bias in the synthetic panel inequality estimates.

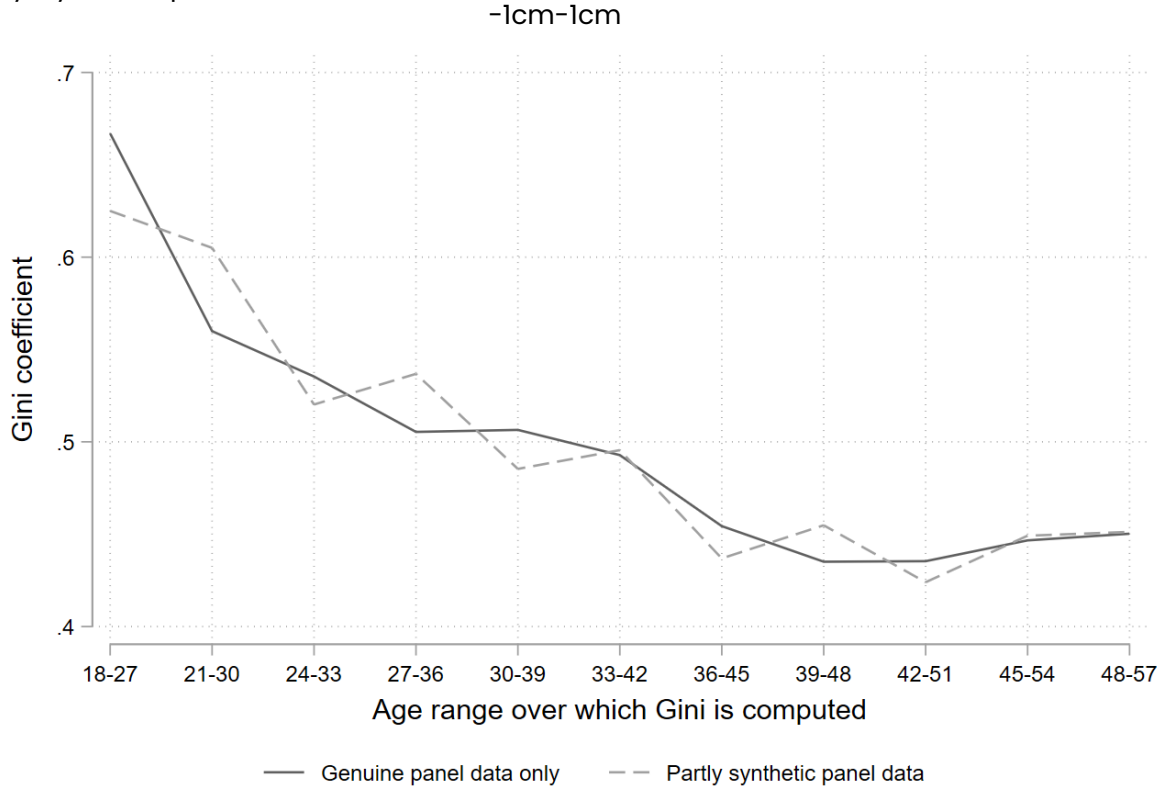
[5] The corollary of this proposition is that a zero standard deviation in the differences would result from a perfect match between observed and imputed earnings.

Figure 3: Evaluating match quality in the construction of the synthetic panel: Standard deviation of the differences between imputed and observed earnings



Notes: This figure reports the standard deviations of the *differences* of observed earnings compared to imputed earnings for the same individuals (solid line). The absolute difference between observed and imputed earnings is computed for each individual where both observed and imputed earnings are available. To do so, a variant of the matching strategy underlying the construction of the synthetic panel is used. In this case, instead of using a "match-on-2, extend-by-2" strategy, the periods on which observations are matched is shifted back one period, and a "match-on-2, extend-by-1" strategy is used. For instance, individuals which are observed at ages 18, 21, 24 and 27 are matched with individuals observed at ages 21, 24, 27 and 30 on variables which are shared at ages 21 and 24. Earnings at age 27 can then be imputed for the younger cohort from the older cohort (i.e. using a "match-on-2, extend-by-1" strategy). However, the younger cohort now has both an *observed* wage, and an *imputed* wage (from the older observation), allowing for a direct comparison between the two. The sample is limited to the NIDS balanced panel and to those aged between 18 and 60 years at the time of their interview. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017.

Figure 4: Gini coefficients of earnings over 9 years, comparing genuine panel estimates with partly-synthetic panel estimates for the same individuals



Notes: The figure reports Gini coefficients, computed using cumulative reported monthly earnings over an approximate 9 year period, across different age ranges. For each individual, earnings are observed at four discreet ages, each 2-3 years apart. Cumulative earnings are inclusive of periods of non-employment with zero earnings. The solid line reports the 9-year earnings Gini coefficient computed using *only genuine panel data* - i.e. where each of the four discreet points at which earnings are observed are unchanged from the original NIDS data. The dashed line reports the 9-year earnings Gini coefficient computed using *partly synthetic panel data* - where three of the four earnings observations are as directly reported by the individual respondent, while the forth earnings observation is imputed using the "match-on-2, extend-by-1" matching strategy underlying the construction of the synthetic panel, as described in this section and in the notes of Figures 3 and 4. Statistics are weighted using the survey design weights, which are corrected for differential panel attrition. The sample is limited to the NIDS balanced panel and to those aged between 18 and 60 years at the time of their interview. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017.

5.3. Sensitivity of lifetime inequality estimates to the matching strategy used in the synthetic panel

The choice of using a "match-on-2, extend-by-2" strategy in the construction of the synthetic panel is justified in Section 4 as a compromise between maximising sample size (which favours a "match-on-1, extend-by-3" approach) and maximising match quality (which favours a "match-on-3, extend-by-1" approach). In Table 5 we evaluate the sensitivity of inequality estimates to this choice.

The leftmost column reports the lifetime inequality estimates using the favoured "match-on-2, extend-by-2" strategy (these are identical to the estimates reported in Table 4), while the middle and rightmost columns report estimates using the other two matching strategies. The number of observations included in the synthetic panel varies: 975 when using the "match-on-3, extend-by-1" approach, 1,728 when using the "match-on-2, extend-by-2" approach, and 2,376 when using the "match-on-1, extend-by-3" approach.

Inequality estimates for the alternative match-and-extend strategies appear to provide upper and lower bounds to our estimates using the "match-on-2, extend-by-2" approach. In most cases, these estimates fall between the two alternatives presented in Table 5. While noting the imprecision inherent in such an exercise, this sensitivity analysis supports our results as reasonable estimates of lifetime inequality in South Africa.

Table 5: Comparing lifetime inequality estimates using different match-and-extend strategies in the construction of a synthetic lifetime panel

	Match 2, extend 2	Match 3, extend 1	Match 1, extend 3
Gini coefficient			
conditional on earnings >0	0.42	0.40	0.47
unconditional on earnings >0	0.42	0.40	0.48
Theil *T* index	0.29	0.26	0.36
Atkinson index (ε=1)	0.44	0.27	0.37
90/10 ratio	11.62	9.23	15.97
90/50 ratio	2.38	2.47	2.94
50/10 ratio	4.89	3.73	5.43
Observations	1 728	975	2 376

Notes: The table reports Gini, Theil *T* and Atkinson inequality indices, and 90/10, 90/50 and 50/10 earnings ratios for individual earnings summed over approximately 36 working-age years, spanning ages 20-56,21-57 or 22-58 in the NIDS life-time synthetic panel. The inequality measures are as described in Table 2. Column 1 reports lifetime inequality measures when using a synthetic panel constructed using the "Match-on-two, extend-by-two" strategy described in Section 4 and used as the preferred strategy throughout the paper. Columns 2 and 3 report sensitivity of these results when using different matching strategies - respectively, "Match-on-three, extend-by-one" and "Match-on-one, extend-by-three" strategies. The "Match-on-three, extend-by-one" strategy uses more time-variant information for matching observations, and therefore may improve match quality but to do so sacrifices the number of observations used. The "Match-on-one, extend-by-three" strategy uses less time-variant information for matching observations, and while this may sacrifice match quality, this is compensated for by a higher number of observations used. Statistics are weighted using the survey design weights, which are adjusted to account for differential panel attrition in the NIDS panel. To facilitate comparisons across time, all monetary figures are deflated using the Stats SA consumer price indices and are calibrated to March 2017.

5.4. Other limitations

It is important to note several other limitations to this exercise which are irresolvable and which introduce unquantifiable noise and/or bias.

The construction of a synthetic panel in the way that we propose limits the number of observations in the analysis to the number of observations in the baseline cohort(s) used to match to older cohorts. In our case, this results is us using an initial cohort of 1,728 17-22 year old youth to estimate the *national* lifetime earnings distribution. There are further difficulties in making claims of representativity: First, substantial panel attrition in the NIDS sample has eroded representivity over time. This is a particular concern for the balanced panel, for which NIDS has not released weights intended to adjust design weights for attrition. Second, since the synthetic panel is a composite dataset drawn from multiple age cohorts in the NIDS panel, it is unclear what the *population of interest* for inference is. This is primarily a conceptual issue: identifying what the population of interest is the logically antecedent to the technical issue of designing weights for the synthetic lifetime panel.

Another major challenge is the inability of the synthetic panel strategy proposed here to deal with large macroeconomic shocks or secular changes. In reality, individuals are affected by shocks and systemic changes which mark sharp discontinuities in the socio-economic evolution across generations. Our match-and-extend strategy is poorly equipped to deal with these changes, since the only input used to impute values from an older cohort is a propensity score matching algorithm run on a limited set of individual variables. A consequence of this is evident in Table 1, where inter-generational educational attainment in the synthetic panel does not reflect that of South African society, but reflects an artificial transferal of the educational profile of the young and relatively highly educated baseline cohort onto older cohorts.

In essence, this strategy is a first attempt at using imperfect data to provide a first estimate of lifetime inequality in a developing country. There is no doubt that improvements to our strategy can be made, just as there are no doubts that some issues will remain ultimately irresolvable.

Conclusions

In the absence of panel data tracking individuals over their lifetimes, inequality in lifetime earnings cannot be estimated directly from observational data. The construction of a synthetic panel spanning the life-course provides one way around this data limitation. In this paper we use existing South African National Income Dynamics Study panel data to construct such a synthetic lifetime panel and provide the first estimates of inequality of lifetime earnings in a developing country.

Consistent with estimates from most developed and developing country contexts, we find that inequality in lifetime earnings is lower than a point-in-time estimate of earnings inequality in South Africa. This difference is likely due to the elimination of age-related earnings disparities in earnings when measured over the life-course compared to point-in-time estimates, which measures the disparity in earnings of individuals who are observed at different ages. However, just as South Africa's cross-sectional earnings inequality is often estimated to be amongst the highest in the world, our estimates for South Africa's lifetime earnings inequality are higher than equivalent measures in most other country contexts (Bowlus and Robin, 2004, 2012; Bönke et al., 2015; Flinn, 2002; Tejada, 2016).

We also find that in South Africa, inequality in cumulative earnings measured over two to nine years is higher than point-in-time estimates. These higher estimates are likely reflective of inequalities in employment dynamics, where high-income earners are more likely to retain stable employment than low-income earners.

This paper provides the first estimates of inequality in lifetime earnings in South Africa, and one of the first estimates in a developing country context. A particular contribution of this paper is that our synthetic panel strategy allows us to compare point-in-time, to medium-term (2-9 years) to lifetime earnings inequality. This allows us to note that, relative to point-in-time estimates, inequality increases over the medium term but decreases over the life-course. This suggests that in the South African context employment and age-earnings dynamics operate over different time-spans and exert opposing pressure on inequality of earnings measured over time.

We hope that our attempt at creating a synthetic lifetime panel in South Africa can be adapted and used by researchers in fields beyond economic inequality. However, several limitations and areas for improvement ought to be flagged: Further research would do well to investigate whether secular trends and large macroeconomic shocks can be more flexibly accounted for than is the case in this first attempt. If doing so proves to be impossible, this limitation will be a persistent problem in the application of synthetic panels to the study of not only inequality but also other phenomena which play out over time.

References

Bassier, I. and Woolard, I. (**2018**). Exclusive growth: Rapidly increasing top incomes amidst low national growth in south africa. Technical report, REDI3x3 Working Paper.

Bönke, T., Corneo, G., and Lüthen, H. (**2015**). Lifetime earnings inequality in germany. *Journal of Labor Economics*, 33(1):171–208.

Bosworth, B., Burtless, G., and Sahm, C. (**2001**). The trend in lifetime earnings inequality and its impact on the distribution of retirement income. *Center for Retirement Research Working Papers*, page 53.

Bowlus, A. J. and Robin, J.-M. (**2004**). Twenty years of rising inequality in us lifetime labour income values. *The Review of Economic Studies*, 71(3):709–742.

Bowlus, A. J. and Robin, J.-M. (**2012**). An international comparison of lifetime inequality: How continental europe resembles north america. *Journal of the European Economic Association*, 10(6):1236–1262.

Brophy, T., Branson, N., Daniels, R., Leibbrandt, M., Mlatsheni, C., and Woolard, I. (**2018**). National income dynamics study panel user manual. Technical Note Release 2018. Version 1., Southern Africa Labour and Development Research Unit.

Corneo, G. (**2015**). Income inequality from a lifetime perspective. *Empirica*, 42(2):225–239.

Fields, G. S. (**2006**). The many facets of economic mobility. In *Inequality, Poverty and Well-being*, pages 123–142. Springer.

Finn, A. and Ranchhod, V. (**2017**). Short-run differences between static and dynamic measures of earnings inequality in south africa. Technical report, REDI3x3 Working Paper 35. Cape Town: The Research Project on Employment ….

Flinn, C. J. (**2002**). Labour market structure and inequality: A comparison of italy and the us. *The Review of Economic Studies*, 69(3):611–645.

Friesen, P. H. and Miller, D. (**1983**). Annual Inequality and Lifetime Inequality*. *The Quarterly Journal of Economics*, 98(1):139–155.

Haider, S. and Solon, G. (**2006**). Life-cycle variation in the association between current and lifetime earnings. *American Economic Review*, 96(4):1308–1320.

Haider, S. J. (**2001**). Earnings instability and earnings inequality of males in the united states: 1967–1991. *Journal of labor Economics*, 19(4):799–836.

Jacques Silber (auth.), J. S. e. (**1999**). *Handbook of Income Inequality Measurement*. Recent Economic Thought Series 71. Springer Netherlands.

Jenkins, S. P. and Van Kerm, P. (**2006**). Trends in income inequality, pro-poor income growth, and income mobility. *Oxford Economic Papers*, 58(3):531–548.

Kerr, A. (**2018**). Job flows, worker flows and churning in south africa. *South African Journal of Economics*, 86:141–166.

Kim, C., Tamborini, C. R., and Sakamoto, A. (**2015**). Field of study in college and lifetime earnings in the united states. *Sociology of Education*, 88(4):320–339.

Kopczuk, W., Saez, E., and Song, J. (**2010**). Earnings inequality and mobility in the united states: evidence from social security data since 1937. *The Quarterly Journal of Economics*, 125(1):91–128.

Leibbrandt, M., Finn, A., and Woolard, I. (**2012**). Describing and decomposing post-apartheid income inequality in south africa. *Development Southern Africa*, 29(1):19–34.

Leibbrandt, M., Ranchhod, V., and Green, P. (**2018**). Taking stock of south african income inequality. Technical report, WIDER Working Paper.

Leibbrandt, M., Woolard, C., and Woolard, I. (**1996**). The Contribution of Income Components to Income Inequality in South Africa. A Decomposable Gini Analysis. Papers 125a, World Bank - Living Standards Measurement.

Leibbrandt, M., Woolard, I., Finn, A., and Argen, J. (**2010**). Trends in south african income distribution and poverty since the fall of apartheid.

Özler, B. (**2007**). Not separate, not equal: Poverty and inequality in post-apartheid south africa. *Economic Development and Cultural Change*, 55(3):487–529.

Paglin, M. (**1975**). The measurement and trend of inequality: A basic revision. *The American Economic Review*, 65(4):598–609.

Sahota, G. S. (**1978**). Theories of personal income distribution: a survey. *Journal of Economic Literature*, 16(1):1–55.

Schotte, S., Zizzamia, R., and Leibbrandt, M. (**2018**). A poverty dynamics approach to social stratification: The south african case. *World Development*, 110:88–103.

Schumpeter, J. A. (**1955**). *Imperialism and Social Classes: Two Essays*. Ludwig von Mises Institute.

Seekings, J. and Nattrass, N. (**2008**). *Class, race, and inequality in South Africa*. Yale University Press.

Sulla, V. and Zikhali, P. (**2018**). Overcoming poverty and inequality in south africa: An assessment of drivers, constraints and opportunities. Technical report, Washington D.C.: World Bank Group.

Tamborini, C. R., Kim, C., and Sakamoto, A. (**2015**). Education and lifetime earnings in the united states. *Demography*, 52(4):1383–1407.

Tejada, M. M. (**2016**). Lifetime inequality measures for an emerging economy: The case of chile. *Labour Economics*, 42:1–15.

Tregenna, F. (**2011**). Earnings inequality and unemployment in south africa. *International Review of Applied Economics*, 25(5):585–598.

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

Zizzamia, R. and Ranchhod, V. (**2019**). Measuring employment volatility in south africa using nids: 2008-2017.

Agence française
de développement
5, rue Roland Barthes
75012 Paris I France
www.afd.fr

What is AFD ?
The Agence Française de Développement (AFD) Group is a public entity which finances, supports and expedites transitions toward a more just and sustainable world. As a French overseas aid platform for sustainable development and investment, we and our partners create shared solutions, with and for the people of the global South.
Active in more than 4,000 projects in the French overseas departments and some 115 countries, our teams strive to promote health, education and gender equality, and are working to protect our common resources – peace, education, health, biodiversity and a stable climate.
It's our way of honoring the commitment France and the French people have made to fulfill the Sustainable Development Goals.
Towards a world in common.

Publication Director Rémy Rioux
Editor-in-Chief Thomas Melonio

Legal deposit 4th quarter 2020
ISSN 2492 - 2846
Graphic design MeMo, Juliegilles, D. Cazeils
Layout AFD
Printed by the AFD reprography service

To browse our publication:
https://www.afd.fr/en/ressources-accueil