Wages and Wage Inequality During the COVID-19 Pandemic in South Africa

By Timothy Köhler and Haroon Bhorat

DPRU Working Paper 202308
October 2023

DPRU
DEVELOPMENT POLICY
RESEARCH UNIT

DPRU

DEVELOPMENT POLICY
RESEARCH UNIT

Wages and Wage Inequality during the COVID-19 Pandemic in South Africa

DEVELOPMENT POLICY RESEARCH UNIT

TIMOTHY KÖHLER

HAROON BHORAT

Working Paper 202308

ISBN 978-1-920633-54-7

October 2023

Abstract
Because the COVID-19 pandemic affected the supply, demand, and nature of work, the implications for wage inequality are ex ante unclear. In South Africa, a country characterised by extreme income inequality driven by wage inequality, these effects are not yet fully understood due to the unavailability of adequate data. This paper makes use of representative and individual-level survey data not available in the public domain provided by Statistics South Africa, to analyse the evolution of the level and nature of wage inequality and its drivers in the country from 2019 to 2022. We first show that missing wage data in the survey is large and non-randomly distributed, justifying imputation. We show that the imputations in the public data are of poor quality and result in an underestimation of wages across the distribution, but parametrically adjusting the raw data for outliers and missing data yields reliable estimates. We find that pre-pandemic wage inequality was extremely high and stable. At the pandemic's onset, real wages mechanically rose primarily due to a composition effect induced by a regressive distribution of job loss. 70 percent of this rise at the mean is explained by this effect, while changes in the returns to characteristics played a relatively muted role. Not considering this former effect leads to misinterpretations of wage dynamics. Composition-controlled indices suggest the pandemic increased wage inequality up to 8 percent or 5 Gini points at its onset, but this was temporary. As the pandemic progressed and employment partially recovered, wage reductions toward pre-pandemic levels stemmed more from lasting changes in the returns to various characteristics than a more similar worker profile, indicative of a persistence of effects on the structure of the labour market.

JEL codes:
D63; J01; J21; J30; J31

Keywords:
COVID-19; wage; inequality; labour market; South Africa; developing country

Acknowledgements:
This is a revised version of one of Timothy Köhler's PhD dissertation chapters. The authors are grateful for constructive feedback on an earlier version of the work from participants of the Economic Society of South Africa's (ESSA) 2023 National PhD conference at the University of Fort Hare, a Research on Socio-Economic Policy (RESEP) seminar in the Department of Economics at Stellenbosch University, a School of Economics seminar at the University of Cape Town, the Third Annual Pretoria-Stellenbosch PhD Workshop in Economics at Stellenbosch University, and Professor Martine Mariotti at the Australian National University.

Corresponding authors
Timothy Köhler (Junior Research Fellow and PhD candidate)
tim.kohler@uct.ac.za

Recommended citation
Köhler, T. and Bhorat, H. (2023). Wages and Wage Inequality during the COVID-19 Pandemic in South Africa. Development Policy Research Unit Working Paper 202308. DPRU, University of Cape Town.

Contents

1. Introduction

A large literature of the labour market effects of the COVID-19 pandemic in South Africa currently exists. However, studies have largely focused on extensive-margin outcomes such as employment, resulting in a dearth of evidence on intensive-margin adjustments. Wages serve as one intensive-margin outcome of particular interest, as well as how they are distributed. At the time of writing, the distributional consequences of the pandemic on wages are not yet fully understood, primarily a consequence of data availability. Importantly, because the impacts of the pandemic have introduced a series of unusual complications in the interpretation of the wage distribution over time, with respect to both compositional and structural dynamics, the implications for wage inequality are *ex ante* unclear. For instance, by affecting the supply, demand, and nature of work, the pandemic may have affected the returns associated with various individual-level characteristics for those who remained employed. The wage distribution is also of course influenced by job loss, which effectively removes workers from the distribution entirely, and the wage inequality implications are dependent on where such an extensive-margin adjustment occurs in the distribution. For instance, because industry-specific economic activity restrictions may have shielded some worker groups at the bottom of the wage distribution, such as agricultural workers, more than other groups further up the distribution, such as hospitality workers, wage inequality may decrease. On the other hand, if 'essential' or 'remote' occupations are concentrated towards the top of the wage distribution – a reasonable conjecture supported by some empirical evidence (Kerr and Thornton, 2020) – and exhibit lower job loss or wage reduction probabilities than their counterparts, wage inequality may increase, or alternatively result in an upward shift of the distribution accompanied by a sustained, unchanged level of inequality. Arguably then, both structural and compositional labour market dynamics ought to be explicitly considered in any analysis of wage inequality during the pandemic.

Such dynamics are particularly important in the South African context which is characterized by extreme levels of income inequality. It is well-documented that the labour market dominates and drives the country's aggregate income inequality, due both to a large share of the population lacking access to labour market incomes (unemployment) and a very unequal distribution of these incomes among the employed. This latter component has, however, been shown to play a more dominant role. As such, a better understanding of wage inequality is critical in aiding one's understanding South Africa's aggregate income inequality. Prior to the pandemic, empirical studies show that wage inequality in the country has remained high and may have even increased during the post-apartheid period, driven by a compression of the bottom half of the distribution – explained by wage setting through collective

bargaining and minimum wage determinations – coupled with a widening of the top half – explained by increased returns to education and demand for non-routine-intense jobs. However, at the time of writing, no empirical studies exist which analyse the effects of the pandemic on the level and nature of wage inequality in the country.

In this paper, we conduct a micro-econometric analysis of the evolution of the level and nature of wage inequality and its drivers during the first two years of the pandemic in South Africa. By doing so, we seek to answer three key research questions. First, what describes the pre-pandemic wage distribution and hence wage inequality in the South African labour market? Second, how did the wage distribution change in response to the pandemic, both at its onset and as it progressed over time? Third, what were the drivers of the change of the wage distribution in response to the pandemic, both at its onset and as it progressed over time? To address these questions, we use nationally representative, individual-level household survey data from Statistics South Africa's (StatsSA) Quarterly Labour Force Survey (QLFS) from 2019 to 2022. Importantly, to avoid significant data quality issues documented in the literature which are attributable to StatsSA's approach to address non-random missing wage values in the survey, we make use the raw, unimputed QLFS wage data provided by StatsSA not available in the public domain to produce reliable estimates of the wage distribution.

This analysis employs a range of techniques categorised into three components. First, by placing a focus on measurement, we examine the quality of the raw, unimputed wage data by making several analytical comparisons to the public domain data and interrogating the quality of the imputations in the latter, and thereafter adopt two parametric techniques to address outliers and impute for non-random missing data using a method which explicitly accounts for the implicit uncertainty which characterises imputations, and thereafter conduct a multitude of diagnostic tests to examine the quality and sensitivity of these imputations. Second, we estimate and analyse trends in real hourly wages across the distribution, as well as several descriptive and normative wage inequality indices which vary in sensitivity to changes in different parts of the distribution, before and after explicitly accounting for the pandemic-induced change in the composition of workers. Third, we conduct decomposition analyses to identify the drivers of the temporal changes in wages from before to after the onset of the pandemic. We examine these drivers both at the mean and across the entire distribution using Oaxaca-Blinder (OB) and Recentered Influence Function (RIF) decomposition, respectively, to isolate the extent changes in wages can be explained by changes in the characteristics of the employed population versus changes in the returns to their characteristics.

This analysis seeks to make several contributions to the literature. First, it is the first to analyse the evolution of the level and nature of wages and wage inequality during the COVID-19 pandemic for the entire employed population in South Africa. While few other studies have made use of alternative wage data[1] collected during the pandemic in the country, the samples used are smaller and more select and the studies themselves focus on either a particular covariate or a limited time period.[2] Second, it provides updated pre-pandemic wage inequality estimates for the country using the most recent and reliable data. Third, it is the first to analyse wage trends using the longest uninterrupted series of raw, unimputed QLFS wage data, and therefore provides an indication of the stability of estimates when each quarterly dataset is appended to one another.[3] Fourth, it further contributes to the empirical evidence on the behaviour of labour markets in developing countries in times of crisis and during COVID-19 in particular.

The structure of the paper is as follows. In Section 2 we synthesise the empirical literature on wage inequality in South Africa prior to the pandemic. In Section 3 we describe the data used as well as the wage data quality adjustments we undertake for the analysis, and present the results from the diagnostic tests undertaken to examine the quality of these adjustments. Thereafter, we outline the methodologies adopted in the latter two components of our analysis in Section 4. We then present the results for these components in Section 5. In Section 6 we conclude.

2. Pre-pandemic wage inequality in South Africa

Income inequality in South Africa has remained stubbornly high in the post-apartheid period. Such persistence serves as a key challenge during this period when the institutional underpinnings of discrimination have and continue to be removed (Wittenberg, 2017). There is a broad consensus in the literature that the labour market continues to dominate and drive the country's aggregate income inequality (Finn et al., 2016; Wittenberg, 2017; Bhorat et al., 2020; Díaz Pabón et al., 2021; Kerr and Wittenberg, 2021; Leibbrandt et al., 2012; 2021; Bhorat et al., 2022; Leibbrandt and Díaz Pabón, 2022). Labour market income has been estimated to account for between 84 and 90 percent of the aggregate per capita household income Gini coefficient (Leibbrandt et al., 2012; Díaz Pabón et al., 2021). This influence is partially because labour market income represents the dominant share of household

[1] The National Income Dynamics Study – Coronavirus Rapid Mobile Survey (NIDS-CRAM), which is a sample- and individual-based longitudinal telephone survey conducted from May 2020 to May 2021 on a much smaller sample (approximately 5 000 to 7 000 adults) compared to the QLFS.

[2] See Hill and Köhler (2020) and Casale and Shepherd (2021) for their analysis of wages by gender, or Ranchhod and Daniels (2021) for their analysis of wages between February and April 2020.

[3] Kerr and Wittenberg (2021) also obtained the unimputed QLFS wage data for their analysis on the union wage premium, however only for 2011 and 2012.

income in the country (Bhorat et al., 2023), while inequality in the labour market is due both a lack of access to labour market incomes (that is, extreme unemployment) as well as the distribution of these incomes amongst those who are employed. Indeed, in this way the labour market can be regarded as notably segmented in that it reproduces the advantage of a minority of high-paid and high-skilled individuals who are employed in secure and well-regulated jobs which are relatively easily obtained, while reproducing the disadvantage of the more vulnerable majority who compete for jobs in a loose labour market among high unemployment which are often characterised as having inadequate job security and benefits (Bhorat et al., 2020; Díaz Pabón et al., 2021). Importantly however, while South Africa's large amount of unemployment (in other words, zero-income earners) explains a large proportion of aggregate income inequality in the country, wage inequality among the employed explains a larger proportion (Leibbrandt et al., 2012; Kerr and Wittenberg, 2021). It is unsurprising then that, like aggregate income inequality, the country has one of the most unequal wage distributions in the world (Díaz Pabón et al., 2021), and hence a better understanding of wage inequality is critical in aiding our understanding of aggregate income inequality.

There is a large literature on the levels, patterns, and determinants of wage inequality in South Africa, most of which makes use of representative household survey data. Overall, there seems to be a consensus that wage inequality has remained high, and may have even increased, during the post-apartheid period. Leibbrandt et al. (2012) estimate a rise in the Gini coefficient from 0.60 in 1993 to 0.64 for 2008, suggestive of a monotonic rise in wage inequality in the post-apartheid period, however the authors can only make use of two comparable cross-sectional datasets and thus cannot account for variation within this period. Wittenberg (2017) addresses this by stacking 29 cross-sectional household surveys to analyse the entire series from 1994 to 2011. Using several inequality measures, they show that wage inequality does appear to have increased over the period. Their estimated Gini coefficient for 1994 is approximately 0.47 compared to 0.55 in 2011. Apart from different time periods, the differences compared to Leibbrandt et al.'s (2012) estimates are at least partially explained by a different analytical sample: Whereas Leibbrandt et al. (2012) estimate a household-level Gini (that is, using household income per capita for people living in households with labour income), Wittenberg (2017) estimates an individual-level Gini using a sample of wage earners. In addition to sample differences, differences in inequality estimates in South Africa are also explained by methodological differences both within and between surveys, making it challenging to draw definitive conclusions on temporal patterns. After taking stock of such differences between 1993 and 2017, Shifa et al. (2023) show that, despite these differences, all datasets consistently measure extremely high levels of income inequality in the international context which appear to have indeed increased during the post-apartheid period.

The Gini coefficient certainly appears to be the dominant measure in the literature. However, when analysing inequality dynamics, the choice of measure matters due to variation in sensitivity to changes in different parts of the wage distribution. Additional measures can also then help shed light on the drivers of changes in inequality. Wittenberg (2017) uses several measures to show that the increase in overall wage inequality from 1994 to 2011 appears driven by a compression of the distribution below the median combined with a widening above it. In other words, wage inequality decreased at the bottom but increased at the top. The compression at the bottom appears driven by a growth of wages at the bottom relative to the middle of the distribution. Extending the period to 2014 and using an alternative method, Wittenberg (2018) similarly finds that the observed increase in mean real wages over this period was driven by an increase (decrease) in inequality in the top (bottom) half of the distribution. These findings are in line with those of Leibbrandt et al. (2012) above as well as Finn and Leibbrandt (2018) who use wage data from StatsSA's labour force surveys in 2000, 2011, and 2014. These latter authors note that between 2000 and 2011, real wage growth across the distribution exhibits a distinct U-shape. Trends beyond 2011 are however less clear due to data quality issues. Evidence of such wage polarisation was however also found by Bhorat et al. (2020) who examine real wage changes between 2000 and 2015, however their data used for the latter period also suffers from the same data quality issues as in Finn and Leibbrandt (2018), discussed in detail later in Section 3.2.

Several studies have also sought to identify and unpack the reasons for the persistence and rise in wage inequality during the post-apartheid period. Wittenberg (2018) discusses how wage setting through collective bargaining and minimum wage determinations may explain the observed compression in the bottom half of the distribution. Finn and Leibbrandt (2018) and Bhorat et al. (2020) both use a distributional decomposition method – one of the methods employed in this paper's analysis – to explain the rise in inequality. In line with Wittenberg's (2018) argument, Bhorat et al.'s (2020) findings suggest that minimum wages may indeed explain the growth at the bottom, while increasing returns to education and non-routine-intense jobs largely explain the growth at the top. For the middle, the authors find that wage growth was undermined by a change in the composition of workers (particularly in mining and manufacturing) as well as reduced returns to routine-intense jobs. Finn and Leibbrandt (2018) also find that growing returns to education, specifically tertiary education, in addition to experience, appear to explain growth at the top end and hence serve as a dominant driver of increasing inequality. However, these findings ought to be interpreted with a degree of caution given the data quality issues mentioned above and discussed below.

Other studies have also shown that the persistence of wage inequality in South Africa can be partially explained by very low intergenerational mobility. Although the correlation between the wages of parents and that of their adult offspring is usually positive in the international literature, South Africa exhibits an extremely strong correlation. Using a representative sample of males aged 20 to 44 years old, Piraino (2015) estimates an intergenerational earnings elasticity for the country of between 0.62 and 0.68; in other words, more than 60 percent of the wage advantage (or disadvantage) of South African fathers is passed on to their sons. This is in line with the notion of the 'Great Gatsby Curve' which suggests that countries with higher levels of inequality have lower levels of intergenerational mobility. Finn et al. (2016) expand on Piraino's (2015) analysis to find estimates of a similar magnitude, but additionally find that immobility is particularly strong at the bottom of the distribution. While this is in line with the international literature, the magnitude is the estimate is unusually high at approximately 0.90. Such strong correlations are often understood to be indicative of unequal opportunities in the labour market, with inherited circumstances playing an influential role in determining outcomes. Concerningly, Piraino (2015) finds that race serves as one of the strongest predictors of mobility – a particularly discouraging finding more than two decades after the end of apartheid.

3. Data

3.1. The Quarterly Labour Force Survey

The analysis in this paper makes use of over three years' worth (or 14 waves) of individual-level household survey data from Statistics South Africa's (StatsSA) Quarterly Labour Force Surveys (QLFS) for all four quarters of 2019, 2020, and 2021 and the first half of 2022. The QLFS is a nationally representative, cross-sectional (with a rotating panel component) household-based sample survey conducted every quarter since 2008 that contains detailed information on a wide array of demographic and socioeconomic characteristics and labour market activities for individuals aged 15 years and older who live in South Africa. The reader is referred to Statistics South Africa (2020a) and Köhler et al. (2023a) for a detailed description of the survey design as well as changes to its mode and sample following the onset of the pandemic in the country in March 2020.

Importantly, the public domain versions of the QLFS data do not include wage data. This data are typically released with a lag in a separate annual publication (the 'Labour Market Dynamics of South Africa'). As is the case with many household surveys, the QLFS exhibits non-negligible rates of item non-response for questions related to earnings, discussed in more detail below. While it is common for statistical agencies to impute or assign values in such cases to avoid non-response bias, a recent literature has highlighted the notably poor quality of the public domain QLFS wage data due to StatsSA's

imputation approach, also discussed in more detail below. To overcome this issue, in this analysis we merge in the raw, unimputed wage data privately provided by StatsSA for each wave and adopt two parametric statistical techniques to address both outliers and missing data, discussed in detail below. Given this paper's focus on wages, the primary sample here is restricted to working-age (15 to 64 years) employed individuals, resulting in a sample of over 180 000 observations in total or 13 000 in the average wave. The wage estimates in this analysis include all forms of wages from labour market activities, including self-employment, and are measured before taxation and deductions. All estimates are weighted using the survey sampling weights and the standard errors are adjusted for the complex survey design through the use of the cluster (the primary sampling unit (PSU) in the case of the QLFS) and strata variables available in the data.

To account for inflation, throughout this analysis we deflate and express the nominal wage data in June 2022 Rands using StatsSA's headline Consumer Price Index data. Additionally, we express wages earned for each hour worked. As discussed in the preceding paper, the QLFS includes several items relating to working hours which vary by a given worker's number of jobs and their "usual" versus "actual" working hours during a reference day or week. The reader is referred to the preceding paper for a detailed discussion of these items. Consistent with our argument in that paper that "actual" working hours is more appropriate than "usual" working hours in the context of the pandemic when various lockdown regulations created or affected the disparity between hours usually and actually worked, here we make use of data on "actual" working hours for a given worker's main job, where main job is defined as the job where a worker usually works the most hours per week, regardless of the number of jobs they have, with one exception. Adopting this approach for furloughed workers (that is, workers who remained employed by reported zero "actual" working hours but a positive wage value in a given period) would result in undefined hourly wage values and hence bias the wage distribution estimates. One option would be to focus exclusively on the actively employed sample (that is, non-zero hour workers), however doing so would exclude a non-negligible share of workers from the wage distribution – a weighted 16 percent of workers in 2020Q2 as shown in the preceding paper. Instead, we retain these workers in the sample but make the explicit assumption that furloughed workers received their "usual" hourly wage (calculated using "usual" working hours data) in a given period, and hence regard being furloughed as a type of paid leave under such circumstances. This approach may result in a degree of measurement error; however, given the absence of more detailed data on wages in the survey, it is arguably amongst the most appropriate of approaches. As a robustness test, we examine the sensitivity of this paper's findings by excluding furloughed workers from the sample.

3.2. Wage data quality adjustments

Our use of the raw, unimputed QLFS wage data provided by StatsSA is important to discuss in brief given the recent debate surrounding the quality of the public release QLFS wage data, which includes imputations, which has played out among labour market researchers in South Africa (Wittenberg, 2017; Kerr and Wittenberg, 2019a; Bhorat et al., 2021; Kerr, 2021; Kerr and Wittenberg, 2021; Köhler et al., 2023b; Köhler, 2023). First, the survey collects data on wages before taxation and deductions from all employees, employers, and own-account workers. These workers are first asked to report their exact wages in South African Rands, and those that do not are then asked to report the bracket or range that their wage falls into. A substantive issue exists in this regard: in the public QLFS wage data from 2010 onwards, StatsSA have included problematic imputations for the wages of workers who did not report them. These include those who neither reported their wage in exact terms nor in a bracket, as well as those who only reported their bracket.[4] Unfortunately, the public release documents do not include an explanation on how these imputations were conducted. In fact, wage imputations are never even mentioned. However, an internal document examined by Kerr and Wittenberg (2021) suggests that StatsSA employed a hot deck imputation method – in which the reported wage of a given respondent or 'donor' is assigned to a given non-respondent with an identical set of observable characteristics – which the authors argue results in imputations of a notably low quality. Specifically, this approach made use of just four variables: gender, race, seven education categories, and three occupation categories. Moreover, StatsSA's approach accounts for bracket responses in a very crude way by making use of only two bracket response categories: less than R6 000 per month and more than R6 000 per month. This strongly contrasts with the survey's 19 possible bracket response categories[5] and can result in very inaccurate imputations. For example, a worker who reported earning between R6 000 and R8 000 per month could be given an imputed wage of any value above R6 000. Unfortunately, the publicly released data does not make it possible to distinguish between the imputed responses and the actual responses. Overall, this suggests that any analysis which makes use of the public QLFS wage data in its current form is erroneous to some degree.

Several studies have highlighted how the use of this public release wage data produces implausible results. Kerr and Wittenberg (2019a) and Kerr (2021) show that these imputations result in unreliable trends in several measures of wage inequality, including the Gini coefficient, the variance of log wages, and trends in five different percentiles. In two unpublished presentations, Khanyile and Kerr

[4] Kerr and Wittenberg's (2021) analysis suggests that, from 2010Q1 to 2012Q2, imputations were made for complete refusals and all bracket responses including refusals and 'don't knows'; however, from 2012Q3 refusals were no longer imputed for.
[5] Excluding the 'don't know' and 'refusal' category.

(2022) compared the unimputed QLFS wage data for a select few years to the public domain data, highlighting the poor quality of the imputations in the latter. Notably, Kerr and Wittenberg (2021) compare estimates from unimputed wage data obtained privately from StatsSA for 2011 and 2012 to public release data for the same periods. The authors find that the imputed wage data produces unreliable results, but that the results appear to be much more reliable when the underlying unimputed data is used. This suggests that although the quality of the imputations done by StatsSA is questionable, the underlying wage data is not. At present, the unimputed data has not been made available in the public domain. Considering the poor quality of the wage data in the public release QLFS, our use of the raw, unimputed wage data resolves the data quality issues pertaining to their imputations discussed above.

To examine the quality of these imputations, we merge the raw, unimputed wage data with the public QLFS wage data.[6] By doing so, we are able to examine the distribution of responses among the employed and how this has varied over time, generate imputation flags to distinguish the imputed from the reported data, and analyse the quality of StatsSA's imputations. Between 2019Q1 and 2020Q4,[7] about 32 percent of all employed in the public QLFS sample have imputed wages, and nearly 40 percent of all wages in the public file are imputed.[8] Of these imputations, most (56 percent) are for cases of completely missing wage data (that is, both exact and bracket responses are missing) while the remainder are for bracket responses. Figure 1 presents the unweighted distribution of wage responses among the employed from 2019Q1 to 2022Q2 using the unimputed data. It should be noted that such a decomposition is not possible with the public QLFS data. The distribution is quite stable over time. Between 45.4 and 53.1 percent of employed individuals in the sample reported an exact wage value in Rand terms, while an additional 18.1 – 22 percent did not report their exact wage but did report a bracket. This latter finding is about consistent with Kerr and Wittenberg's (2021) analysis of the 2011 and 2012 unimputed data which showed bracket responses comprised 20 – 23 percent of employed individuals. Together, this implies that the average wave tends to have non-missing wage data of some kind for nearly two-thirds of all workers, with missing wage data then for over one-third of workers. While the survey instruments of course differ in design, this missing data rate is not dissimilar from the

[6] It should be noted that, since 2020Q2, the item in the QLFS instrument which asks respondent workers to report the exact value of their wage included an explicit instruction to enumerators to enter the value zero if the respondent did not state their wage. This instruction is not included in the instrument for any waves prior to 2020Q2. The implication of this instruction is that researchers would be unable to distinguish true zero values from non-responses, which is particularly relevant at the pandemic's onset when many workers became furloughed. However, all waves of raw data provided by StatsSA do not include any zero values, which implies that such values were recoded as missing.

[7] This merge can only be done using the public QLFS wage data for 2019 and 2020 given that the public data beyond this period was not yet made available at the time of writing.

[8] These two shares are not equivalent because the public QLFS wage data does not include imputations for all workers with missing wage data.

US Current Population Survey (CPS) which typically contains missing earnings data for around 30 percent of the employed (Bollinger and Hirsch, 2006; Kerr and Wittenberg, 2021).

Figure 1: Distribution of wage responses among the employed in the QLFS, 2019Q1 – 2022Q2

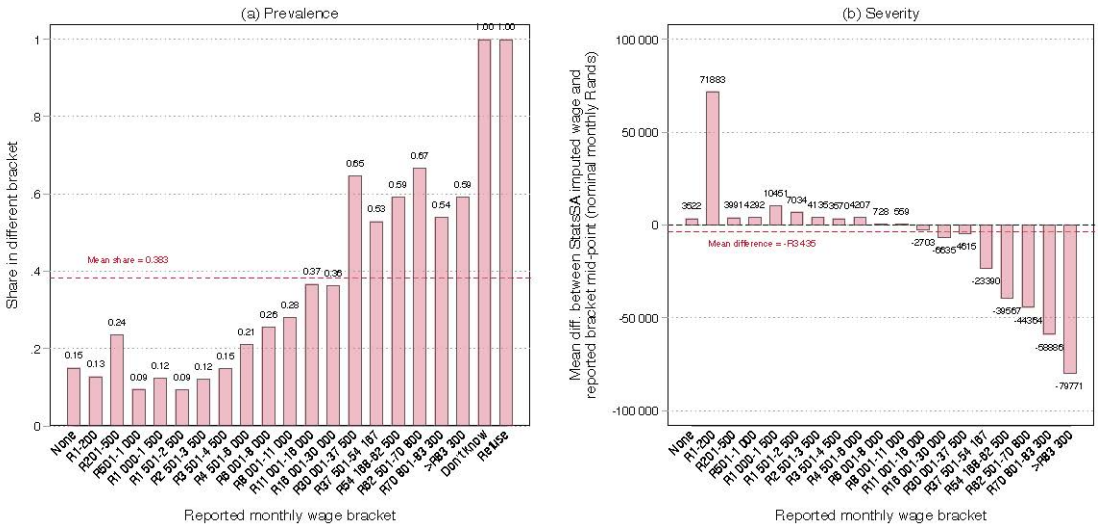
Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Unweighted estimates presented. DK = Don't know bracket responses.

Expectedly, this missing data does not appear to be distributed randomly but instead is associated with several observable covariates. As shown in Table A1 in the appendix, relative to workers who only reported bracket data or neither bracket nor exact data, the average worker who did report their exact wage value is statistically significantly younger, have fewer years of education, more likely to be female, African/Black, work in the informal sector, live in a rural area, work in the private sector, and not be a trade union member. Such differences are also reflected in a more conditional environment, such as in Table A2 which presents pooled Linear Probability Model estimates of the predictors of non-response. Given that all of these characteristics are associated with lower wages in the South African labour market, this indicates that wage non-response is non-random and is likely concentrated towards the top of the wage distribution, which is consistent with the literature (Wittenberg, 2017).

The quality of the public QLFS wage imputations is apparent when analysing the imputation values among those who did not report their exact wage value but did report the bracket within which their

wage lies. For such responders, one would expect a reasonable imputation procedure to bound their imputed values within their reported bracket. However, in two unpublished presentations, Khanyile and Kerr (2022) showed that the imputations for bracket responders are largely outside the brackets individuals actually reported. Building on their work, in Figure 2 we present two measures of 'inaccuracy' of these imputations for 2020Q1: the 'Prevalence' measure in panel (a) considers the share of imputations which fall into a different bracket other than the reported one, and the 'Severity' measure in panel (b) considers, among those in a different bracket, the absolute difference between the imputed value and the bracket mid-point.

Figure 2: Inaccuracy measures of public QLFS wage imputations, 2020Q1.



Author's own calculations. Source: QLFS 2020Q1 (Statistics South Africa, 2020a).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed who reported bracket wage data. Unweighted estimates presented. Panel (a) considers the share of imputations in the public QLFS data which fall into a different bracket other than the reported one. Panel (b) considers, among those in a different bracket, the absolute difference between the imputed value and the bracket mid-point.

The two panels indeed strongly suggest that the public QLFS wage imputations are of a poor quality. Excluding don't knows and refusals, 38.3 percent of imputations for bracket responders are outside the bracket which the respondent reported their wage lies in.[9] This is consistent with the analyses by Khanyile and Kerr (2022) referred to above. Notably, this share varies considerably across the reported bracket distribution, with larger shares among higher-wage workers. For example, while 12 percent of imputations for workers who reported earning between R2 501 and R3 500 per month are outside this range, the equivalent share for workers who reported earning between R62 501 and R70

[9] The figure also shows that StatsSA also imputed wages for all workers who refused to report their bracket or reported that they did not know their wage. This contrasts Kerr and Wittenberg's (2021) analysis which showed that refusals were no longer imputed for in 2012Q3, which suggests that StatsSA changed their imputation approach sometime thereafter.

800 per month is 67 percent. The degree these imputations lie outside reported brackets is not negligible. As shown in panel (b), excluding bracket don't knows and refusals, the average imputation outside a reported bracket is R3 435 lower than the bracket mid-point. Notably, there are very severely inaccurate imputations for the R1 – R200 bracket, within which the average imputed monthly wage outside the bracket is nearly R72 000 larger than the bracket mid-point.[10] The data is also indicative of a growing discrepancy towards the top of the bracket distribution.

The unimputed wage data by itself is, of course, also not immune to non-random item non-response. To prepare the unimputed data for analysis, we follow Wittenberg (2017) and Kerr and Wittenberg (2019b) and adjust the data to (i) identify outliers and (ii) address missing values. We discuss these two approaches in detail below.

3.2.1. Outlier detection

We employ a studentised regression residual approach to identify outlying wage values and recode them as missing. While there are several outlier detection algorithms available, the studentised regression residual approach is advantageous in that it addresses outliers in both tails of the distribution, not only at the top end. This approach entails estimating an expanded Mincerian wage regression of the logarithm of monthly wages[11] on a vector of observable covariates using Ordinary Least Squares (OLS), predicting the residuals, and flagging observations with large residuals as outliers. Conceptually then, outlying wage values are considered as those which deviate significantly from what would be expected as implies by the parameters in a model of the determinants of wages. Here, the vector of observable covariates includes the usual Mincerian covariates – years of education and potential experience[12] (and its squared term) (Mincer, 1974; Limieux, 2006; Patrinos, 2016) – as well as age, sex, racial population group, province, an urban indicator, marital status, main industry and occupation, a public sector indicator, a formal sector indicator, a trade union membership indicator, and

[10] Strikingly, for one observation in 2020Q1 who did not report their exact wage but reported a bracket of R1 – 200 per month, StatsSA imputed a monthly wage of R404 434. This seems like a very implausible wage given this discrepancy, but additionally given that a worker of this set of characteristics (a 53-year old woman with seven years of education working in a sales and services occupation in the informal sector in rural Limpopo) is not associated with such a high wage. As an illustration, an expanded Mincerian regression model on the observed (exact) wage data for the wave predicts a wage of just R279 per month for this worker.

[11] There are no workers in any period in the dataset who exhibit zero monthly wages, so taking the logarithmic transformation does not result in a smaller or more select sample.

[12] Experience is not observed in the data, so potential experience is derived as

survey wave fixed effects.[13][14] As shown in panel (a) of Figure 3, the residuals are concentrated around zero and appear randomly distributed across the fitted values, which suggests that both linearity and homoscedasticity hold. However, a few larger residuals are evident, but making a judgement on their magnitude is difficult because residuals depend on the unit of measurement. Additionally, points of high leverage tend to be associated with smaller residuals. Studentised residuals address these problems by adjusting each residual by an estimate of its standard deviation. The studentised residual for individual $i - r_i$ – is defined as follows:

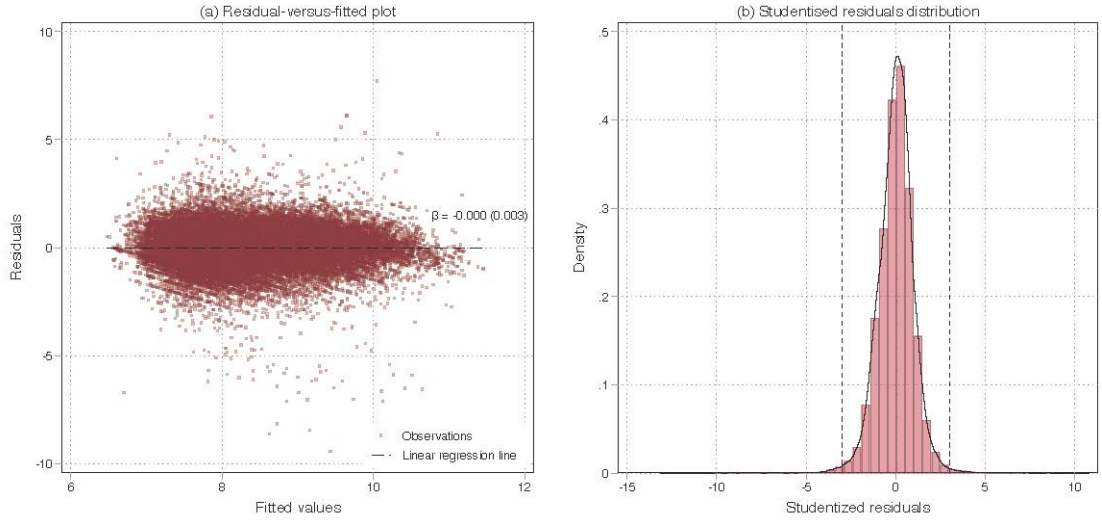
$$r_i = \frac{\varepsilon_i}{\sqrt{s^2_{(i)}(1 - h_i)}} \tag{1}$$

where ε_i is the unstandardised residual, s_i^2 is the estimated variance of the residual with the i^{th} observation removed, and h_i is the leverage. As expressed by Wittenberg (2017), r_i can be interpreted as a t-statistic for testing the significance of a dumour variable equal to one in a given observation and zero elsewhere, so such a variable effectively absorbs the observation and remove its influence on the other coefficients in the model. The distribution of the studentised residuals is presented in panel (b) in Figure 3. Following Stevens (1984), outliers are defined as observations with absolute studentised residuals in excess of three, which then detects about one percent (n = 894) of reported exact wages as outliers. These outliers are evenly distributed across survey waves. These wages are recoded as missing and then imputed for along with other observations with missing wage data using the approach discussed below.

[13] Of course, the model will only include observations with non-missing data on all the included covariates. The extent of missing data for one covariate in particular – trade union membership status – is non-negligible at about 16 percent of worker observations in the period. To retain these workers in the sample, they are assigned the wave-specific mean of trade union membership status and a binary 'missing trade union membership status' variable is included as an additional control to flag these observations.

[14] Although the original specification by Mincer (1974) proposed modelling wages parsimoniously as a linear function of years of schooling and a quadratic function of years of potential experience, it is common in the contemporary literature to expand the model to include additional covariates of interest.

Figure 3: Residuals-versus-fitted-values plot and the studentised residuals distribution



Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Residuals and fitted values obtained by estimating an expanded Mincerian wage regression of the logarithm of monthly wages on a vector of observable covariates using OLS. Model is unweighted. Vertical lines in panel (b) correspond to a value of three in absolute terms.

3.2.2. Multiple imputation

Wittenberg (2017) discusses how there are two broad approaches for dealing with missing data: reweighting non-missing values to account for missing ones or imputing for the missing data. While several methods are available, in our analysis here we employ a multiple imputation (MI) approach. First proposed by Rubin (1976), MI is now considered as one of the most effective methods for addressing item non-response (Daniels, 2023). The approach is similar to stochastic imputation which first imputes a single value, parametrically or non-parametrically, and then adds a random error term to the predicted value. One key issue with stochastic imputation is that subsequent statistical analysis treats the imputed value as the true value, even though it is the sum of the true value and some measurement error. In other words, the imputed value does not reflect any of the uncertainty implicit in the imputation process. MI is advantageous in that it repeats the imputation process multiple times to produce multiple values of what the true data might have been. Appropriate point estimates and standard errors are then obtained using Rubin's (1976) rules, which state that standard complete-data techniques should be used to estimate the variance of estimators *within* all of the complete datasets while accounting for differences in estimates *between* datasets. Formally, Rubin's (1976) rules are defined as follows, following the exposition by Daniels (2023). For the estimated parameter θ, the mean is simply computed as:

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m \tag{2}$$

where $\hat{\theta}_m$ is a complete-data estimate for $m = 1, \ldots, M$ imputations. The within and between components of the variance, $\bar{W}_M$ and B_M respectively, are:

$$\bar{W}_M = \frac{1}{M}\sum_{m=1}^{M} W_m \tag{3}$$

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M} \left(\hat{\theta}_m - \bar{\theta}_M\right)^2 \tag{4}$$

The total variance, T_M, can then be obtained by combining (3) and (4) as follows:

$$T_M = \bar{W}_M + \frac{M+1}{M} B_M \tag{5}$$

Confidence intervals can be calculated, and significance tests conducted, using a t distribution, $(\theta - \bar{\theta}_M) T_M^{-1/2} \sim t_v$ with $v = (M-1)\left(1 + \frac{1}{M+1}\frac{\bar{W}_M}{B_M}\right)^2$ degrees of freedom.

In our analysis, MI is used to impute exact wage values for workers who (i) neither reported their exact wage nor their bracket (including here those who reported the 'refusal or 'don't know' bracket), (ii) only reported their bracket, and (iii) were identified as outliers as discussed in the previous section. Imputations are not generated for those who reported exact wage values and were not detected as outliers. Because the missing wage data in the QLFS has a monotone pattern – that is, if bracket wage data is missing then exact wage data is missing – due to the questionnaire's skip logic, imputations here are generated by specifying a sequence of independent univariate conditional imputation methods. Separately for each wave and following Wittenberg (2017), we first multiply impute a bracket for those in group (i) or (iii) by estimating an ordered logit model on a vector of observable covariates, and thereafter multiply impute log monthly wages based on the imputed bracket and the same vector of observable covariates using predictive mean matching (PMM) with 10 nearest neighbours.[15] For observations in group (ii), the imputation process of course skips the first step and proceeds with

[15] PMM entails regressing log monthly wages on the (imputed or reported) bracket and the vector of observable covariates, and then matches observations with missing and non-missing wage data using their predicted log monthly wage. In other words, the *actual* wage from an observation with non-missing wage data is imputed for an observation with missing wage data but a similar *predicted* wage. As such, this process is defined even for workers with missing exact wage data provided they have non-missing explanatory variable data.

multiply imputing log monthly wages as described above. This process is repeated iteratively to arrive at 10 imputations, and we set the seed to ensure reproducibility. A similar approach was followed by Kerr and Wittenberg (2019b) in their generation of the PALMS[16] dataset – a compilation of individual-level microdata from household surveys conducted between 1993 and 2019 in South Africa. Following Van Buuren et al. (1999), the selection of observable covariates to be included is based on those which are required in the complete data model of interest, those which appear to determine missingness (see the relevant Linear Probability Model estimates presented in Table A1), and those which explain a considerable amount of the variance of log monthly wages. These are included in both imputation models following the recommended procedure (Rubin, 1987), and include age, sex, racial population group, years of education, potential experience (and its squared term), province, an urban indicator, marital status, main industry and occupation, a public sector indicator, a formal sector indicator, frequency of wage payments, and a trade union membership indicator.[17]

Table 1 presents information on the sample sizes, extent of missing data, and number of imputations for both bracket and exact value responses over the period. In the pooled sample, 52 percent of workers do not report exact wage data, while 32 percent do not report bracket wage data, which implies that nearly 40 percent of those that do not report exact wage data do report bracket wage data. In other words, the average wave tends to have non-missing wage data (either exact or bracket responses) for nearly two-thirds of workers, with missing wage data then for over one-third of workers. This is, expectedly, consistent with the estimates presented in Figure 1. As previously discussed, the extent of missing data is relatively constant over time. Finally, as shown in columns (6) and (10), imputations were successfully made for nearly all observations with missing bracket data and missing exact wage data (97 percent in both cases). As with the missing data rate, these imputation rates were also relatively constant over time.

[16] The Post-Apartheid Labour Market Series.
[17] Observations with missing trade union membership status data here are treated similarly as with the outlier detection model.

Table 1: Sample size, item non-response, and imputation information, 2019Q1 – 2022Q2

	Total employed (n)	Brackets				Exact values			
		Missing data, incl. DK/Refuse (n)	Missing data rate (%)	Imputations (n)	Imputation rate (%)	Missing data (n)	Missing data rate (%)	Imputations (n)	Imputation rate (%)
	(1)	(3)	(4) = (3)/(1)	(5)	(6) = (5)/(3)	(7)	(8) = (7)/(1)	(9)	(10) = (9)/(7)
2019Q1	17 490	5 464	31.2	5 243	96.0	9 072	51.9	8 722	96.1
2019Q2	17 414	5 372	30.8	5 136	95.6	9 208	52.9	8 817	95.8
2019Q3	17 597	5 251	29.8	5 042	96.0	9 068	51.5	8 708	96.0
2019Q4	17 422	5 078	29.1	4 913	96.8	8 876	50.9	8 575	96.6
2020Q1	17 044	4 976	29.2	4 805	96.6	8 646	50.7	8 340	96.5
2020Q2	10 001	2 879	28.8	2 795	97.1	4 686	46.9	4 526	96.6
2020Q3	10 464	3 456	33.0	3 370	97.5	5 457	52.2	5 320	97.5
2020Q4	11 008	3 677	33.4	3 574	97.2	5 841	53.1	5 684	97.3
2021Q1	10 200	3 590	35.2	3 498	97.4	5 508	54.0	5 364	97.4
2021Q2	11 827	4 211	35.6	4 097	97.3	6 397	54.1	6 227	97.3
2021Q3	8 938	3 170	35.5	3 130	98.7	4 810	53.8	4 726	98.3
2021Q4	8 041	2 804	34.9	2 736	97.6	4 309	53.6	4 210	97.7
2022Q1	10 448	3 479	33.3	3 397	97.6	5 549	53.1	5 405	97.4
2022Q2	12 947	4 608	35.6	4 468	97.0	7 068	54.6	6 859	97.0
Total	180 841	58 015	32.1	56 204	96.9	94 495	52.3	91 483	96.8

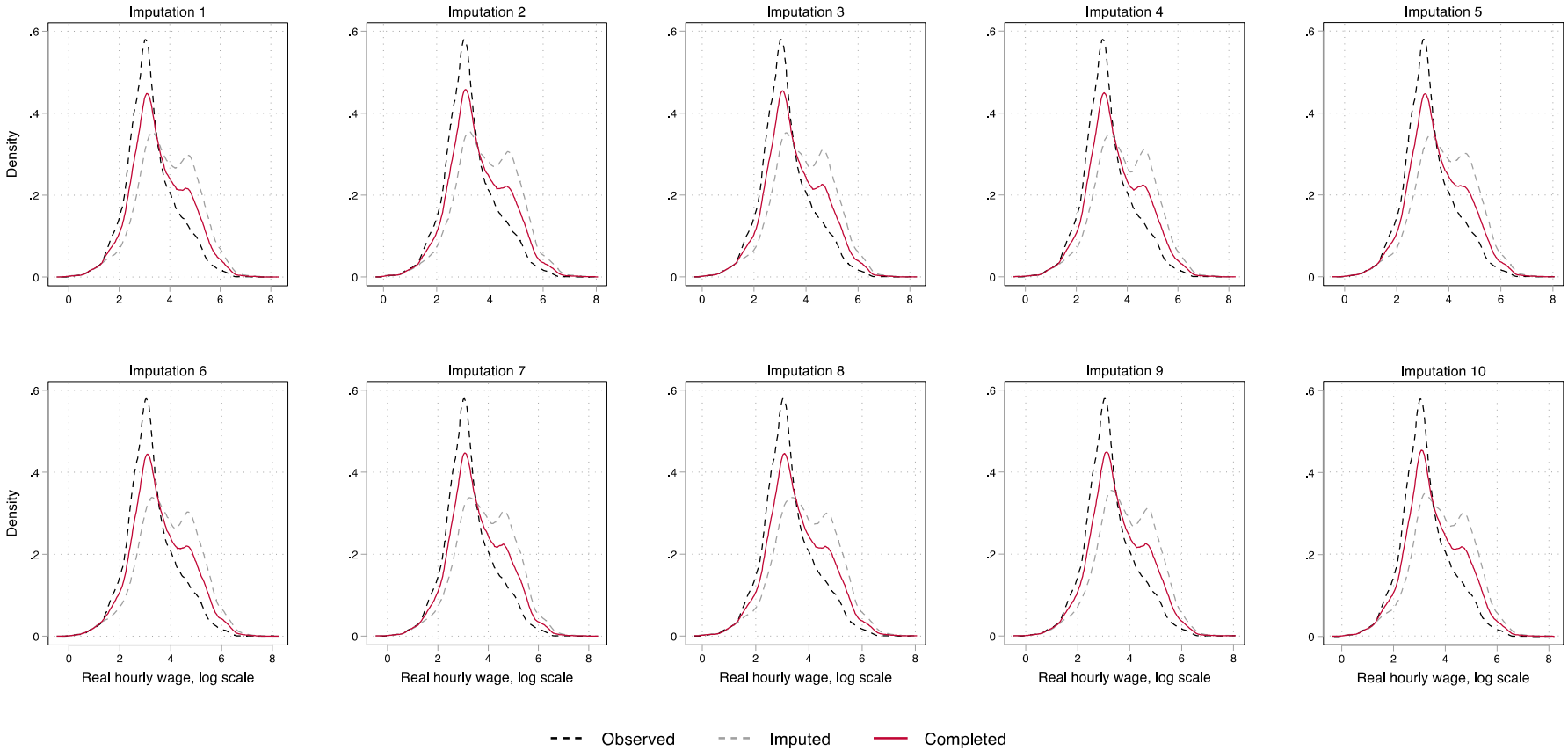
Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Horizontal dashed line refers to the onset of the COVID-19 pandemic in South Africa. Number of observations in columns (5) and (12) in a given wave refers to the minimum number of observations for which wage data was imputed for among the 10 imputation iterations.

we conduct several diagnostic tests to assess the quality of the imputations, including comparing wage distributions across imputation iterations, examining how the distributions of complete (the sum of observed and imputed values) and imputed values only compare to the distributions of observed values only and the distribution implied by the public QLFS data, comparing the complete distributions across different types of responders (for example, those who only reported bracket responses to those who neither reported exact nor bracket values), and analysing how estimates of the complete distribution vary by the number of imputations and alternative imputation model specifications.

First, following Abayomi et al. (2008), using data for 2020Q1,[18] we estimate and plot kernel density estimates of the wage distributions separately using the observed, imputed, and complete (the sum of the observed and imputed) data for each of the 10 imputations. We plot these estimates in Figure 4. Differences between the observed and imputed distributions are expected here given the assumption that the wage data is missing not at random (MNAR) – that is, the probability of reporting wages varies across the wage distribution, and in particular tends to have an inverse relationship with wages (Wittenberg, 2017). As such, it may be expected that the distribution of the imputed data is located more rightwards relative to the observed data. Indeed, this appears to be the case for each imputation iteration. The imputed data distributions are all towards the right of the observed distributions, and the p-values of Kolmogorov–Smirnov tests of the equality of these distributions are all close to 0.000, implying significantly different distributions. Moreover, the imputed data distributions all exhibit a similar shape to one another and are not indicative of unreasonable wage values.

[18] The relevant distributions for other survey waves are not shown for brevity, however they all exhibit similar characteristics to the 2020Q1 data.

Figure 4: Diagnostic plot of real hourly wage distributions by sample and imputation iteration, 2020Q1



Author's own calculations. Source: 2020Q1 (Statistics South Africa, 2020a).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Unweighted estimates presented. Wages adjusted for inflation and expressed in June 2022 Rands. Observed = non-imputed wage data only; Imputed = imputed wage data only; Completed = combination of observed and imputed data.

Table 2: Mean and median real hourly wage estimates by dataset, 2019Q1 – 2022Q2

	QLFS public wage data					Exact wage data					Incl. imputations				
	n	Mean		Median		n	Mean		Median		n	Mean		Median	
	(1)	(2)		(3)		(4)	(5)		(6)		(7)	(8)		(9)	
2019Q1	14 571	68.61	(2.56)	24.45	(0.27)	8 419	48.39	(1.00)	23.52	(0.28)	17 124	73.43	(1.46)	32.96	(0.62)
2019Q2	14 499	66.51	(2.24)	24.01	(0.25)	8 205	48.52	(1.05)	23.10	(0.25)	17 014	74.79	(1.51)	32.51	(0.71)
2019Q3	14 640	66.76	(2.17)	24.40	(0.25)	8 529	48.16	(1.02)	22.89	(0.25)	17 220	74.43	(1.58)	33.02	(0.56)
2019Q4	14 421	89.19	(22.66)	24.45	(0.24)	8 548	47.75	(1.29)	23.04	(0.24)	17 105	75.43	(1.98)	32.91	(0.57)
2020Q1	14 103	66.18	(2.08)	24.36	(0.24)	8 399	45.57	(1.03)	22.74	(0.23)	16 721	71.72	(1.58)	32.49	(0.55)
2020Q2	8 430	101.78	(26.02)	28.94	(0.44)	5 314	56.78	(1.51)	26.04	(0.41)	9 841	86.85	(2.74)	37.02	(1.02)
2020Q3	5 479	63.31	(3.12)	24.99	(0.40)	5 008	49.80	(1.14)	24.62	(0.39)	10 327	74.91	(1.61)	36.26	(1.03)
2020Q4	8 912	60.45	(2.25)	26.43	(0.38)	5 168	50.02	(1.40)	24.31	(0.38)	10 851	74.49	(1.89)	34.61	(0.90)
2021Q1	.	.	.	.	.	4 692	50.25	(1.36)	23.96	(0.36)	10 056	77.08	(2.10)	35.16	(1.05)
2021Q2	.	.	.	.	.	5 430	47.98	(1.44)	23.68	(0.32)	11 655	73.43	(2.10)	33.33	(0.79)
2021Q3	.	.	.	.	.	4 127	48.14	(1.73)	23.30	(0.40)	8 854	76.67	(3.24)	34.15	(1.09)
2021Q4	.	.	.	.	.	3 732	44.08	(1.67)	23.04	(0.40)	7 942	67.86	(2.58)	31.90	(0.91)
2022Q1	.	.	.	.	.	4 900	45.28	(1.51)	23.85	(0.33)	10 304	65.89	(2.01)	31.66	(0.97)
2022Q2	.	.	.	.	.	5 878	51.20	(1.80)	23.49	(0.28)	12 738	70.76	(1.70)	31.94	(0.69)
Total	95 055	73.12	(4.45)	25.37	(0.11)	86 349	48.74	(0.36)	23.58	(0.08)	177 752	74.07	(1.03)	33.34	(0.43)

Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Horizontal dashed line refers to the onset of the COVID-19 pandemic in South Africa. Wages adjusted for inflation and expressed in June 2022 Rands. Estimates weighted using sampling weights. Standard errors are adjusted for the complex survey design and are presented in parentheses. QLFS public wage data only available for 2019 and 2020 at the time of writing.

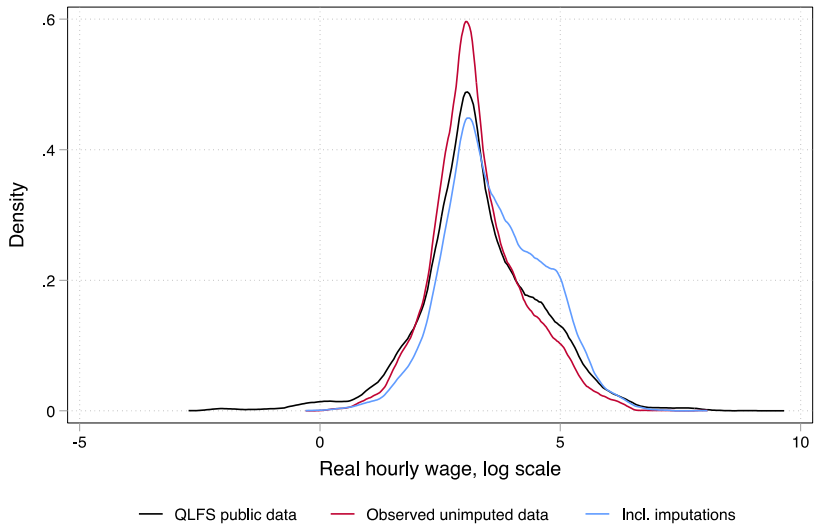
Another simple diagnostic for the MI approach entails the examination of mean and median estimates using the data before and after the inclusion of the imputed data – that is, the observed and complete data. We present the relevant estimates in Table 2, along with the equivalent estimates using the public QLFS wage data for comparison.[19] Overall, the table suggests that this study's approach results in a larger sample and more precise estimates in a given period which are less volatile over time. This is indicative that the imputed values obtained from the MI model are reasonable. As shown in columns (7) to (9), using the pooled complete data results in an estimated mean of R74 and median of R33, which appear to be relatively constant over time.[20] These estimates are notably higher than those obtained using only the observed data, as shown in columns (4) to (6), which is expected given the MNAR nature of the wage data and the relationship between wages and the probability of reporting wages discussed above. The larger standard errors of the complete relative to the observed data estimates are also expected given that the MI procedure explicitly incorporates additional uncertainty into the estimates. Using the public QLFS wage data, as shown in columns (1) to (3), the smaller sample sizes imply that StatsSA's approach imputed wages for a smaller share of workers in the sample, which may be the reason behind the inflated standard errors. Relative to the complete case, while the data results in a lower median of R25 but a similar mean of R73, the latter appears to be influenced by a subset of extremely high values in 2019Q4. This outcome is presumably a consequence of StatsSA's imputation approach given that it is not evident in either the complete or observed data. Disregarding the 2019Q4 data however reduces the mean to between R60 and R69 over the period.

The characteristics described above suggest that the use of either the public QLFS wage data or the observed data alone results in an underestimation of wages. This appears to be the case not only when considering mean and median values but also across the entire distribution, as presented in Figure 5. Relative to the complete data distribution which includes the imputations here, both the distributions of the observed data only and the public QLFS data are positioned towards the left. At the bottom, the public QLFS data exhibits 10th and 25th percentiles of the lowest values (R7 and R14, respectively), as reflected by the distribution's longer bottom tail. These estimates are lower than the equivalent estimates using either the observed data (R9 and R15) or the complete data (R12 and R19). Towards the top of the distribution, the complete data percentiles also exceed those of both the observed data and public QLFS data.[21] Analysing the distributions of imputed values using the public QLFS data versus

[19] For the public QLFS wage data, real hourly wages only for 2019Q1 to 2020Q4 could be estimated given that the QLFS public wage data for both 2021 and 2022 were not yet made available in the public domain at the time of writing.
[20] Apart from a spike at the onset of the pandemic in 2020Q2, which is evident in all datasets here and as such cannot be a consequence of the imputation process. An examination of this outcome is deferred to the detailed discussion in Section 5.
[21] The 75th and 90th percentile values for each distribution are as follows, respectively: R88 and R169 for the complete data; R58 and R142 for the public QLFS data; and R46 and R104 for the observed data.

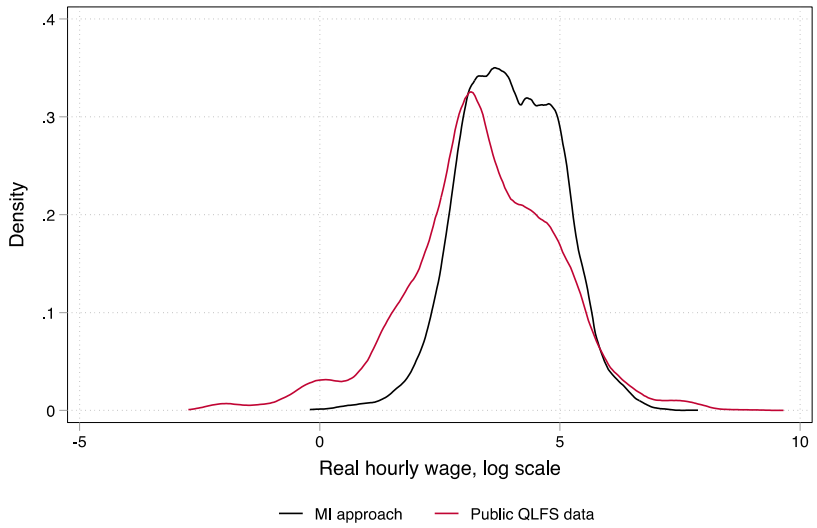
those obtained through the MI approach here, presented in Figure 6, reveals that while both approaches exhibit similar means of R94 and R96 respectively, StatsSA has imputed more extreme wage values at both tails of the distribution. The distribution's 1st and 99th percentiles are approximately R0.37 and R1 182 respectively, compared to those of R5 and R534 in the alternative distribution, resulting in a 50 percent lower median of R29.

Figure 5: Real hourly wage distributions by dataset, 2020Q1



Author's own calculations. Source: 2020Q1 (Statistics South Africa, 2020a).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates weighted using sampling weights. Wages adjusted for inflation and expressed in June 2022 Rands.

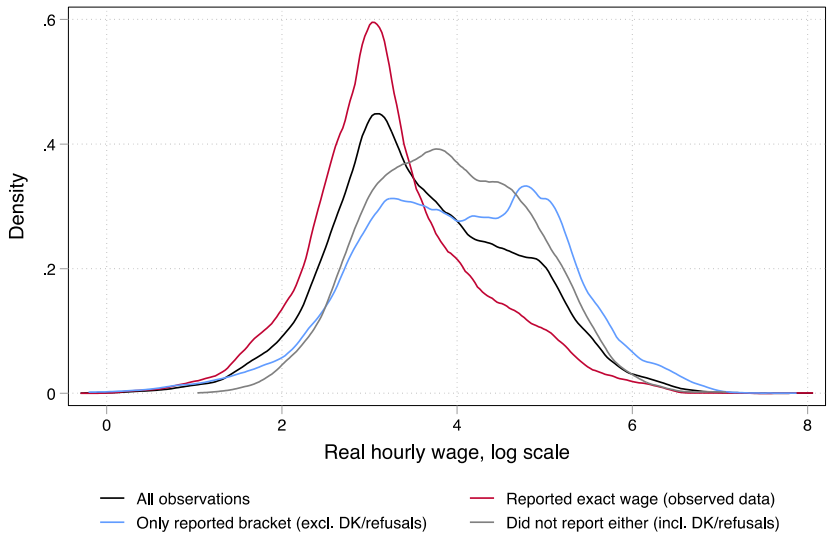
Figure 6: Distributions of imputed real hourly wages by dataset, 2020Q1



Author's own calculations. Source: 2020Q1 (Statistics South Africa, 2020a).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed whose wages were imputed. Estimates weighted using sampling weights. Wages expressed in June 2022 Rands.

Next, we disaggregate the distribution presented in Figure 6 to examine the subsets of the multiply imputed wage data; that is, the imputed data for workers who did not report their exact wage but did report their bracket and those who did not report either. These distributions are presented in Figure 7 along with the distributions of the complete and observed (exact wage) data for comparison. It is apparent that the imputed wage distributions of both sources of missing data are relatively similar. The distribution for those who reported bracket information (excluding 'don't know' and refusal' responses) exhibits a mean and median of R100 and R57 respectively, compared to R92 and R60 for those who neither reported their exact wage nor their bracket (including 'don't know' and refusal' responses), in other words 'complete missings'. Both distributions also exhibit similar degrees of positive skewness, however the kurtosis of the distribution for bracket responders is higher at 51.7 compared to 33.2 for 'complete missings', as reflected by the former distribution's longer tails. Importantly, the figure shows that the densities of imputations lie to the right of the exact data distribution. This is consistent with Daniels' (2023) findings, who uses an MI approach on alternative labour force survey data for South Africa in the late 1990s and early 2000s, and is expected given the positive correlation between wages and the probability of non-response discussed prior. Finally, while higher wages are typically imputed for, the figure makes it clear that both the minimum and maximum wage values in the complete distribution stem from the exact and not from imputed draws.

Figure 7: Real hourly wage distributions by type of wage response, 2020Q1



Author's own calculations. Source: 2020Q1 (Statistics South Africa, 2020a).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates weighted using sampling weights. Wages expressed in June 2022 Rands.

We next analyse the stability of the mean and median wage estimates by varying the number of imputations. We consider the cases of two, five, and 20 imputations, compared to our primary estimates which makes use of 10 imputations. These estimates are presented in Table 3. Again, following Rubin's (1976) rules, the estimates for a given individual are computed as the mean of their multiply imputed values. It is clear that, regardless of the number of imputations here, both the mean and median estimates are almost identical across the number of imputations. This holds both within a given survey wave and over time, both before and after the onset of the pandemic. The sudden rise in estimates at the onset of the pandemic in 2020Q2 is also evident. While the discussion of this increase is deferred to Section 5, the observation that it occurs regardless of the number of imputations again suggests that it is not a consequence of the imputation process. The precision of the estimates, as reflected by the standard errors, also do not vary considerably as the number of imputations increase. In other words, the relationship between the number of imputations and inference does not appear to be strong in the data here. As such, it can be concluded that the estimates here are very stable across varied number of imputations and, in line with Daniels (2023), stability of multiply imputed income data can be achieved with as little as two multiple imputations.

Table 3: Mean and median real hourly wage estimates by number of imputations, 2019Q1 – 2022Q2

	m=2				m=5				m=20			
	Mean		Median		Mean		Median		Mean		Median	
	(1)		(2)		(3)		(4)		(5)		(6)	
2019Q1	73.93	(1.49)	33.06	(0.59)	73.49	(1.63)	33.10	(0.58)	73.33	(1.43)	32.89	(0.61)
2019Q2	73.70	(1.36)	32.64	(0.93)	74.34	(1.60)	32.46	(0.71)	75.17	(1.76)	32.27	(0.63)
2019Q3	73.72	(1.47)	33.02	(0.55)	73.99	(1.46)	33.02	(0.55)	74.28	(1.46)	33.02	(0.55)
2019Q4	76.40	(2.15)	32.91	(0.58)	75.55	(2.01)	32.91	(0.57)	76.24	(2.35)	32.91	(0.57)
2020Q1	72.83	(2.33)	32.49	(0.55)	72.05	(1.90)	32.49	(0.55)	71.93	(1.60)	32.49	(0.55)
2020Q2	87.70	(2.94)	37.61	(0.94)	87.48	(2.85)	37.38	(1.00)	86.91	(2.92)	37.32	(0.99)
2020Q3	74.48	(1.91)	36.25	(1.07)	75.01	(1.86)	36.37	(1.04)	75.25	(1.85)	36.31	(1.01)
2020Q4	75.32	(2.25)	34.54	(0.79)	74.77	(1.85)	34.63	(0.84)	74.74	(1.89)	34.64	(0.98)
2021Q1	76.36	(1.80)	34.99	(1.03)	76.22	(1.77)	34.86	(0.92)	76.56	(1.97)	35.11	(0.96)
2021Q2	73.40	(1.62)	33.39	(0.87)	74.15	(2.30)	33.50	(0.80)	73.73	(2.01)	33.36	(0.82)
2021Q3	76.34	(2.53)	34.24	(1.00)	76.56	(2.66)	34.35	(1.14)	76.40	(2.67)	34.38	(1.15)
2021Q4	67.34	(2.45)	31.76	(0.84)	67.61	(2.87)	31.85	(0.88)	67.19	(2.44)	31.98	(0.93)
2022Q1	64.99	(2.52)	31.30	(0.74)	66.06	(2.78)	31.64	(1.08)	65.92	(2.35)	31.59	(0.86)
2022Q2	72.17	(2.45)	32.35	(0.81)	71.59	(1.99)	32.29	(0.68)	71.85	(1.96)	32.23	(0.70)
Total	74.14	(1.10)	33.35	(0.43)	74.15	(1.03)	33.36	(0.43)	74.20	(1.03)	33.34	(0.43)

Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).

Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Horizontal dashed line refers to the onset of the COVID-19 pandemic in South Africa. Wages adjusted for inflation and expressed in June 2022 Rands. Estimates weighted using sampling weights. Standard errors are adjusted for the complex survey design and are presented in parentheses.

As a final diagnostic for the MI approach adopted here, we test the sensitivity of estimates to a range of varied specifications of the prediction models in the imputation algorithm. Four models are developed for this purpose. First, an intentionally-misspecified model is estimated which only includes gender and province as predictors in the observable covariate vector. Second, a model which only includes covariates which predict missing wage data is estimated. As shown in Table A1, this includes wage frequency, age (and its squared term), gender, years of education, racial population group, province, main industry and occupation, and urban and public sector employment indicators. Third, a model which only includes Mincerian wage function covariates is estimated, which includes years of education and potential experience (and its squared term). Finally, the fourth model is estimated using covariates from both a Mincerian wage function and those which predict missing wage data. As described by Daniels (2023), the first model serves as a baseline to provide insight into the importance of covariate misspecification in the imputation algorithm; the second model generates imputations which are "uncongenial" in nature – that is, the imputation model differs from the intended complete analysis model; the third model then generates imputations which are more "congenial" to analysing wages even though covariates which are associated with the response process are absent; while the fourth model, a-priori, is most similar to the main imputation specification described earlier in this section and hence is treated as first-best as it conforms to the recommendations of Van Buuren et al. (1999).[22]

Estimates of mean and median wages for each survey wave obtained using the four alternative multiple imputation algorithm specifications are presented in Table 4. For a given wave, the estimates from the second model – which only include covariates which predict missingness – and the fourth model – which also includes these covariates along with those from a typical Mincerian wage function which are more "congenial" to analysing wages – are similar to one another as well as to those obtained using the main imputation specification in this paper's main analysis. In contrast, both the intentionally-misspecified model and that which only includes Mincerian wage function covariates produce notably smaller mean and median wage estimates. Given that the only difference between models 2 and 4 is the inclusion of potential experience and its squared term, it should be noted that many of the covariates which predict missingness also explain a non-negligible share of the variation in wages in an 'expanded' Mincerian wage function, such as main occupation and industry. These results then suggest that covariate selection based on explaining the response process, as well as the outcome variable of interest (wages here), are particularly crucial for drawing plausible wage values using the data here. This is consistent

[22] Three covariates are not included in this model's specification but are in the main imputation specification: marital status, a formal sector indicator, and a trade union membership indicator. The reason for this discrepancy is that they neither predict missingness nor are typical Mincerian covariates, but they are required in the complete data model of interest.

with Daniels (2023) who notes that specifying MI algorithms using covariates which explain the response process alone is suboptimal. Additionally, it should be noted that the rise in wages in 2020Q2 is evident regardless of the algorithm specification, which strongly suggests that the rise is not a consequence of the imputation procedure.

Table 4: Mean and median real hourly wage estimates across alternative imputation model specifications, 2019Q1 – 2022Q2

	(i) Intentionally misspecified				(ii) Predict missingness only				(iii) Mincerian				(iv) Mincerian + predict missingness			
	Mean		Median		Mean		Median		Mean		Median		Mean		Median	
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
2019Q1	66.53	(1.58)	28.57	(0.44)	74.46	(1.53)	33.16	(0.63)	68.30	(1.39)	30.50	(0.49)	74.79	(1.61)	33.19	(0.66)
2019Q2	67.20	(1.64)	28.02	(0.44)	76.13	(1.64)	32.78	(0.74)	69.67	(1.46)	29.99	(0.57)	75.91	(1.73)	32.57	(0.83)
2019Q3	67.07	(1.60)	28.58	(0.39)	75.22	(1.68)	33.02	(0.55)	69.48	(1.48)	30.44	(0.49)	75.16	(1.56)	33.02	(0.55)
2019Q4	65.79	(2.29)	28.29	(0.53)	76.36	(2.27)	32.91	(0.57)	70.62	(2.08)	30.20	(0.52)	76.26	(2.50)	32.91	(0.56)
2020Q1	64.19	(1.44)	28.32	(0.49)	72.33	(1.55)	32.49	(0.55)	68.58	(1.77)	30.49	(0.54)	72.49	(1.59)	32.49	(0.55)
2020Q2	76.40	(2.39)	32.46	(0.71)	86.91	(2.80)	37.33	(0.96)	80.77	(2.46)	34.77	(0.79)	86.79	(2.58)	37.20	(1.07)
2020Q3	66.49	(2.10)	30.61	(0.94)	76.32	(1.91)	36.66	(0.96)	69.88	(2.01)	32.98	(0.99)	76.12	(1.66)	36.60	(1.05)
2020Q4	65.92	(1.84)	29.37	(0.68)	75.27	(1.77)	34.68	(0.93)	68.05	(1.72)	31.90	(0.63)	75.71	(1.85)	34.67	(0.87)
2021Q1	67.75	(1.83)	29.53	(0.87)	77.04	(2.05)	35.44	(1.14)	70.71	(2.13)	31.62	(0.74)	76.82	(2.01)	35.01	(0.94)
2021Q2	62.11	(1.81)	27.71	(0.47)	74.43	(2.45)	33.10	(0.83)	66.79	(2.01)	30.59	(0.68)	73.75	(2.15)	33.13	(0.79)
2021Q3	64.04	(2.40)	28.52	(0.74)	76.28	(2.53)	34.15	(1.08)	69.01	(2.38)	30.57	(0.70)	76.38	(2.46)	34.04	(1.11)
2021Q4	56.76	(1.97)	27.31	(0.61)	69.01	(3.26)	31.76	(0.95)	61.18	(2.11)	29.68	(0.85)	69.42	(2.48)	31.96	(0.95)
2022Q1	56.73	(1.82)	26.65	(0.48)	66.88	(2.73)	31.51	(1.03)	59.52	(1.91)	29.06	(0.71)	66.40	(2.17)	31.18	(0.91)
2022Q2	62.88	(2.37)	27.22	(0.48)	73.45	(2.07)	32.25	(0.71)	66.08	(1.79)	29.27	(0.50)	73.68	(1.90)	32.42	(0.67)
Total	65.00	(0.79)	28.57	(0.27)	74.96	(1.05)	33.37	(0.44)	68.46	(0.87)	30.70	(0.34)	74.93	(1.05)	33.36	(0.45)

Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b). Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Horizontal dashed line refers to the onset of the COVID-19 pandemic in South Africa. Wages adjusted for inflation and expressed in June 2022 Rands. Estimates weighted using sampling weights. Standard errors are adjusted for the complex survey design and are presented in parentheses.

4. Methodology

4.1. Aggregate trends in wages and wage inequality

This paper's analysis is structured in two components. In the first, we estimate and analyse trends in real hourly wages and several commonly-used wage inequality indices for the weighted sample of workers in the South African labour market from the pre-pandemic baseline period (2019Q1 – 2020Q1) through to after the onset of the pandemic (2020Q2) and during its first two years (up to and inclusive of 2022Q2). For the analysis of wages, we adopt a distributional analysis by explicitly examining cross-sectional estimates and temporal changes across the entire wage distribution. For the analysis of wage inequality, to gain a comprehensive understanding of wage dispersion across the entire distribution, we make use of several measures given that they vary in their sensitivity to changes in different parts of the distribution, discussed below. We use both descriptive and normative measures; that is, ones which are calculated using only mathematical formulae and ones which are additionally derived from a social welfare function, respectively. We primarily use measures which are relative and not absolute in nature. The former are generally preferred to the latter because they have the advantage of being scale invariant – that is, if all wages were multiplied by one positive scalar, the relative inequality measure will remain unchanged – which is a desirable property because it ensures that the inequality measure is insensitive to the units in which wages is measured (Allison, 1978; Shorrocks, 1984; Sen, 1997; Atkinson and Brandolini, 2010; Shifa and Ranchhod, 2019). Specifically, we estimate and examine the following indices: the Gini coefficient, the Atkinson index, Theil's T index, as well as various percentile ratios and quantile shares. While all of these measures try to describe the distribution of wages in some way, they vary in the level of importance placed at different parts of the distribution. These are described in more detail below.

4.1.1. The Gini coefficient

The Gini coefficient is one of the most commonly used measures of inequality. Formally, it can be calculated as per Shifa and Ranchhod (2019) as follows:

$$Gini = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{|y_i - y_j|}{2N^2\mu} \quad (6)$$

where y_i and y_j represent the wages of worker i and j, respectively, μ the mean wage, and N the size of the population of workers. The coefficient ranges between 0 and 1 with higher values indicating higher inequality. It can be visualised as a Lorenz curve – a graphical representation of a distribution (in this case, wages) which plots the cumulative share of wages earned by the poorest x percent of a population for all possible values of x. A 'curve' of a 45-degree line represents perfect equality; that is, when wages are shared equally among all individuals, however these curves tend to exhibit a convex shape given the generally unequal distribution of wages (the poorest x percent of a population earn less than x percent of total income). The Gini coefficient can then be calculated as the area between the Lorenz curve and the 45-degree line as a proportion of the total area under the 45-degree line. An advantage of the coefficient is that it uses data from the entire distribution to generate a summary statistic, but it places greater weight on the middle of the distribution. However, an important limitation of the index is that a similar coefficient between different groups or time periods need not imply similar distributions.[23]

4.1.2. The Atkinson index

The Gini coefficient is a descriptive measure, implying that its calculation does not entail the incorporation of an explicit social welfare function. Atkinson (1970) however argued that such a measure does assume some implicit value judgement because they are used in policymaking processes. To allow for a measure which explicitly incorporates a social welfare function, Atkinson (1970) proposed the Atkinson class of inequality measures which are one of the most commonly-used normative inequality measures in the literature. These measures reflect the welfare loss to a society due to inequality (Shifa and Ranchhod, 2019). They do so by explicitly including an inequality aversion parameter which can vary between zero and infinity, with greater values implying that a society more heavily weights a given transfer towards the lower end of the distribution relative to an equivalent transfer towards the top. The measures are computed as follows:

$$Atkinson(\varepsilon) = 1 - \left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i}{\mu}\right)^{(1-\varepsilon)}\right]^{\frac{1}{(1-\varepsilon)}} \quad (7)$$

[23] As an illustration, in a context where half a population earning zero wages while the other half shall all wages equally, compared to a context where 75 percent of a population earns 25 percent of wages shared equally while the remaining quarter earn 75 percent of wages shared equally, it would be reasonable to consider the latter context as more equal than the former because half of the population in the former earn nothing. However, both contexts will exhibit the same Gini coefficient of 0.5.

where y_i, μ, and N take on the same definitions as in equation (1). ε represents the inequality aversion parameter. Although the choice of which is somewhat arbitrary, the most-commonly used values in the literature are 0.5, 1, 1.5, or 2 (Sen, 1997; Atkinson and Brandolini, 2010; Shifa and Ranchhod, 2019). In our analysis here we incorporate $\varepsilon = 1$, which makes the measure sensitive to changes in inequality at the bottom of the distribution. Like the Gini coefficient, values of the measure vary between 0 and 1, regardless of the choice of parameter, and is interpretated with respect to an equal income distribution. In the case of wages, a value of 0.80 implies that 80 percent of all wages is 'wasted' due to inequality, or alternatively, just 20 percent of all wages is needed to achieve a level of social welfare equivalent to one with an equal wage distribution.

4.1.3. Theil T index

While the Atkinson class of measures have the advantage that they make the social welfare function in a given context explicit, they can result in different rankings of income distributions depending on the choice of the inequality aversion parameter (Cowell, 2011; McGregor et al., 2019). An alternative set of measures – the class of Generalised Entropy (GE) measures – overcomes this disadvantage. At their core, these measures are based on ratios of incomes to the mean income, and as such can be useful in understanding which part of the distribution drives an observed change in inequality. These measures are calculated as follows:

$$GE(\alpha) = \frac{1}{\alpha(\alpha - 1)}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i}{\mu}\right) - 1\right] \tag{8}$$

where y_i, μ, and N again take on the same definitions as in equation (1). α is a parameter which represents the weight given to inequality at different parts of the distribution. The greater a positive α value, the more sensitive the GE measure is to changes in inequality at the top of the distribution. This parameter can be of any real value, although the most commonly-used values are 0, 1, or 2 (Shifa and Ranchhod, 2019). When $\alpha = 1$, the measure is often referred to as the Theil T index – one of the most popular GE measures – which is sensitive to changes in inequality at the top of the distribution, unlike the Atkinson index which is sensitive to changes at the bottom when $\varepsilon = 1$ and the Gini coefficient which is sensitive to changes in the middle. In our analysis here we employ this specific GE measure. Unlike the Gini and Atkinson index, the values of the GE measures themselves are not restricted to vary between 0 and 1 but instead vary between 0 and infinity, with higher values again representing higher levels of inequality.

4.1.4. Percentile ratios and quantile shares

While the above measures make use of the entire income distribution, percentile ratios and quantile shares focus only on specific parts of the distribution. Percentile ratios serve as a comparison of incomes at different parts of the distribution, and quantile shares serve as a measure of income concentration in a particular part of the distribution. The fact that they only make use of two parts of the distribution at a time can be regarded as both a disadvantage – because they do not reflect information from the entire distribution – as well as an advantage – because they are transparent about which part of the distribution is driving any observed changed in a summary inequality measure. However, one can overcome the aforementioned disadvantage by simply computing a multitude of ratios and shares. To calculate them, we first order the population of workers in a given period from poorest to richest and then categorise them into specific quantile groups (for example, quintiles or deciles). Thereafter, to calculate quantile shares we estimate the proportion of total wages that accrue to each quantile group in a given period. In our analysis here, we estimates quantile shares for the bottom 50 percent, the middle 40 percent (workers who earn between the 30^{th} and 70^{th} percentile of the distribution), the top 10 percent, and the top 5 percent of workers. To calculate percentile ratios, we simply divide the wage at a particular percentile (for instance, the 90^{th} percentile or p90) by the wage at another percentile (for instance, the 10^{th} percentile or p10). As such, values can range between zero and infinity and the higher the value, the greater the level of inequality. We calculate ratios for p90 to p10 (the 90/10 ratio), p90 to p50 (the 90/50 ratio), and p50 to p10 (the 50/10 ratio) to analyse wage disparities between the upper end and the bottom, the upper end and the middle, and the middle and the lower end of the wage distribution, respectively.

4.2. Decomposition analysis of changes in wages at the mean and across the distribution

The second component of this paper's analysis consists of an examination of the drivers of the temporal changes in wages from before to after the onset of the pandemic in South Africa. In other words, we seek to decompose wage inequality over time, and hence this component is dynamic in nature. We do so both at the mean and across the entire distribution, and in doing so we seek to understand how and to what extent such changes can be explained by the relative contributions of changes in the characteristics of the employed population and changes in the returns to these characteristics. We first conduct the analysis at the mean using a twofold Oaxaca-Blinder (OB) decomposition, introduced by Oaxaca (1973) and Blinder (1973), and thereafter employ Recentered Influence Function (RIF) regression (also known as unconditional quantile regression) and decomposition for the distributional

analysis, which was introduced by Firpo et al. (2009) and expanded by Fortin et al. (2011) as a means of generalising the OB decomposition to any unconditional quantile of an outcome distribution. We outline both these procedures in more detail below.

Regarding the twofold OB decomposition, we are interested in comparing wages between two time periods, $t \in (1,2)$. OB decomposition is related to the earlier developed Kitagawa decomposition and has the same objective, however OB decomposition is more general and is only identical to Kitagawa decomposition under very specific circumstances (Oaxaca and Sierminska, 2023). Following Oaxaca (1973) and Blinder (1973), assuming wages can be expressed as a linear function of observable and unobservable covariates:

$$wage_{it} = \boldsymbol{X_{it}}\beta_t + \varepsilon_{it}, \text{ for } t \in (1,2) \tag{9}$$

A model which pools data for both periods can then simply be expressed as follows:

$$wage_i = \boldsymbol{X_i}\beta + \varepsilon_i \tag{10}$$

If an indicator variable $T = 0$ for $t = 1$ and $T = 1$ for $t = 2$, then the following can represent the difference in wages at the mean across periods:

$$\begin{gathered} E[wage_i|T = 1] - E[wage_i|T = 0] \\ = E[\boldsymbol{X_i}|T = 1]'(\beta_2 - \beta) + E[\boldsymbol{X_i}|T = 0]'(\beta - \beta_1) + (E[\boldsymbol{X_i}|T = 1] - E[\boldsymbol{X_i}|T = 0])'\beta \end{gathered} \tag{11}$$

Equation (9) can be estimated as follows, where horizontal bar accents represent sample means:

$$\begin{gathered} \overline{wage}_{i2} - \overline{wage}_{i1} = \left[\overline{\boldsymbol{X}}'_{\boldsymbol{i2}}\left(\hat{\beta}_2 - \hat{\beta}\right) + \overline{\boldsymbol{X}}'_{\boldsymbol{i1}}\left(\hat{\beta} - \hat{\beta}_1\right)\right] + (\overline{\boldsymbol{X}}'_{\boldsymbol{i2}} - \overline{\boldsymbol{X}}'_{\boldsymbol{i1}})\hat{\beta} \\ = \hat{\Delta}_S^{\mu} + \hat{\Delta}_X^{\mu} \end{gathered} \tag{12}$$

The first term in equation (10), $\hat{\Delta}_S^{\mu}$, is referred to the estimated wage structure effect which speaks to the relative contribution of changes in the returns to characteristics in the vector $\boldsymbol{X_{it}}$ to temporal wage changes at the mean μ, while the second term, $\hat{\Delta}_X^{\mu}$, is referred to the estimated composition effect which speaks to the relative contribution of changes in the characteristics of the employed (again those in the vector $\boldsymbol{X_{it}}$) to temporal wage changes at the mean. These components are sometimes alternatively referred to as the 'price' and 'quantity' components, respectively. While the estimated

coefficients can be used to obtain overall structure and composition effects for all covariates, it also provides estimates of the structure and composition effects for each covariate j as follows:

$$\hat{\Delta}_S^{\mu} = \sum_{j=1}^{k} \bar{X}'_{2,j}\left(\hat{\beta}_{2,j} - \hat{\beta}_j\right) + \bar{X}'_{1,j}\left(\hat{\beta}_j - \hat{\beta}_{1,j}\right) \tag{13}$$

$$\hat{\Delta}_X^{\mu} = \sum_{j=1}^{k} \left(\bar{X}'_{2,j} - \bar{X}'_{1,j}\right)\hat{\beta}_j \tag{14}$$

The Recentered Influence Function (RIF) decomposition approach to analyse these temporal changes beyond the mean and across the entire distribution operates in a similar way to the OB decomposition. The exception is that the outcome variable in a RIF regression is the RIF of any functional of the outcome instead of the outcome itself. These functionals may be specific quantiles of the outcome distribution, or specific distributional statistics such as the Gini coefficient or percentile ratios. If f is the functional of the distribution, then $\hat{\Delta}_S^{\mu}$ and $\hat{\Delta}_X^{\mu}$ in the case of the mean μ can be expressed in the case of f as follows, similar to equations (11) and (12):

$$\hat{\Delta}_S^{f} = \sum_{j=1}^{k} \hat{\Delta}_{S,j}^{f} = \sum_{j=1}^{k} \bar{X}'_{2,j}\left(\hat{\beta}_{2,j}^{f} - \hat{\beta}_j^{f}\right) + \bar{X}'_{1,j}\left(\hat{\beta}_j^{f} - \hat{\beta}_{1,j}^{f}\right) \tag{15}$$

$$\hat{\Delta}_X^{f} = \sum_{j=1}^{k} \hat{\Delta}_{X,j}^{f} = \sum_{j=1}^{k} \left(\bar{X}'_{2,j} - \bar{X}'_{1,j}\right)\hat{\beta}_j^{f} \tag{16}$$

A comparison of equations (13) and (14) in the OB decomposition case to (15) and (16) in the RIF case makes it clear that the latter is identical to the former when the functional f is the mean μ, as discussed by Ferreira et al. (2017).

In our analysis, we begin with the overall and detailed OB decomposition of the logarithm of real hourly wages at the mean, and thereafter conduct overall and detailed RIF decompositions along the percentiles of the distribution. We analyse three periods of interest: the pre-pandemic baseline (2019Q2) to the onset of the pandemic (2020Q2), the pre-pandemic baseline to one year after the pandemic's onset (2021Q2), and the pre-pandemic baseline to two years after the pandemic's onset

(2022Q2). We follow Finn and Leibbrandt (2018) and Bhorat et al.'s (2020) choice of covariates[24] to consider the relative contributions of the following possible drivers: age, race, sex, urbanisation, education (years of schooling), industry, experience (included its squared term), unionisation, and sector of employment (public versus private). We further expand from those included in these studies by additionally including occupation, formality of employment, and province, bringing the total number of drivers considered to 12.[25] Lastly, it should be noted that in this component of the analysis we continue to make use of the multiply imputed wage data described in Section 3.2.2.

5. Results

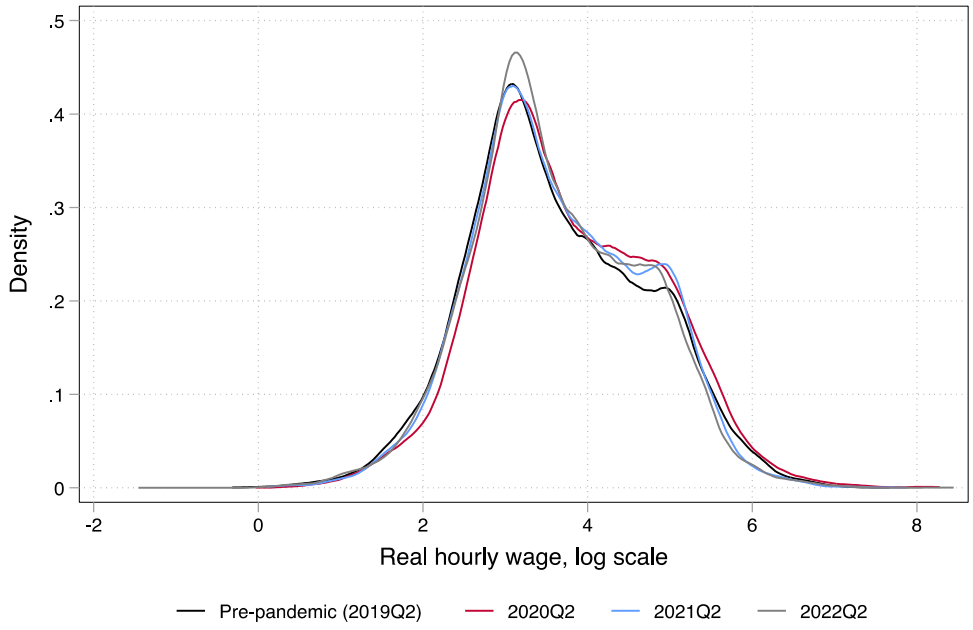
5.1. Aggregate trends in wages and wage inequality

In this section we present the results from our analysis of the trends of real hourly wages and wage inequality from the pre-pandemic baseline period through to after the onset of the pandemic and during its first two years. To begin, in Figure 8 we present kernel density estimates of the real hourly wage distributions over the period. To control for seasonality, we present the quarter 2 distributions for each year. Overall, we observe a clear rightwards but transient shift in the distribution at the onset of the pandemic accompanied, however, by a very marginal change in the shape of the distribution. This shift is reflected by variation in the mean wage (other points of the distribution are explored later). Prior to the pandemic, the estimated mean wage was R74.79 (s.e.[26] = R1.51) per hour worked, or R12 754.60 (s.e. = R248.67) per month. At the onset of the pandemic, these estimates increased to R86.85 (s.e. = R2.74) and R13 387.74 (s.e. = R360.48) respectively, with each difference being statistically significant by at least the 5 percent level. These represent substantially large real year-on-year increases of 16 and 5 percent, respectively. Two-sample Kolmogorov–Smirnov (KS) tests indicate that these distributions are statistically different from each other (p = 0.000). One year later in 2021Q2, the distribution returned to a more similar shape compared to the pre-pandemic period, as reflected by the similar means of real hourly and monthly wages of R73.43 (s.e. = R2.10) and R12 521.46 (s.e. = R360.14), respectively. KS test results do not suggest the 2021Q2 and pre-pandemic distributions are statistically different from one another. These estimates remained relatively stable through to the end of the period another year later in 2022Q2, with marginally lower mean real hourly and monthly wages of R70.76 (s.e. = R1.70) and R11

[24] With the exception that we do not include Bhorat et al.'s (2020) five task content variables coded using an alternative dataset.
[25] Firpo et al. (2018) note that, with respect to categorical variables, the contribution of a given covariate to the wage structure effect for both OB and RIF decomposition is sensitive to the choice of the reference group. Unfortunately, the authors also show that there is no simple solution to this problem. For transparency, the reference groups for categorical variables in the analysis here are as follows: youth (aged 15-34 years), men, those not married or living together with a partner, rural areas, self-reported Black/African individuals, the agriculture industry group at the one-digit level, the managers occupation group at the one-digit level, union non-membership, the private sector, and the informal sector.
[26] The estimated standard error.

826.42 276.458 (s.e. = R276.46), respectively, however these estimates are not statistically different from their pre-pandemic equivalents. Considering the shapes of the distributions, they are all similarly positively skewed with skewness coefficients ranging between 0.16 and 0.23 over the period, however the 2020Q2 distribution exhibits the highest degree thereof. Kurtosis coefficients are relatively constant with coefficients ranging between 2.67 and 2.83, apart from the 2022Q2 distribution which exhibits a marginally higher coefficient of 2.90, as reflected by the longer bottom tail. The variances are largely unchanged and vary between 1.06 and 1.16, suggestive of a little to no change in wage inequality among the employed over the period.

Figure 8: Kernel density estimates of the real hourly wage distribution, 2019 – 2022

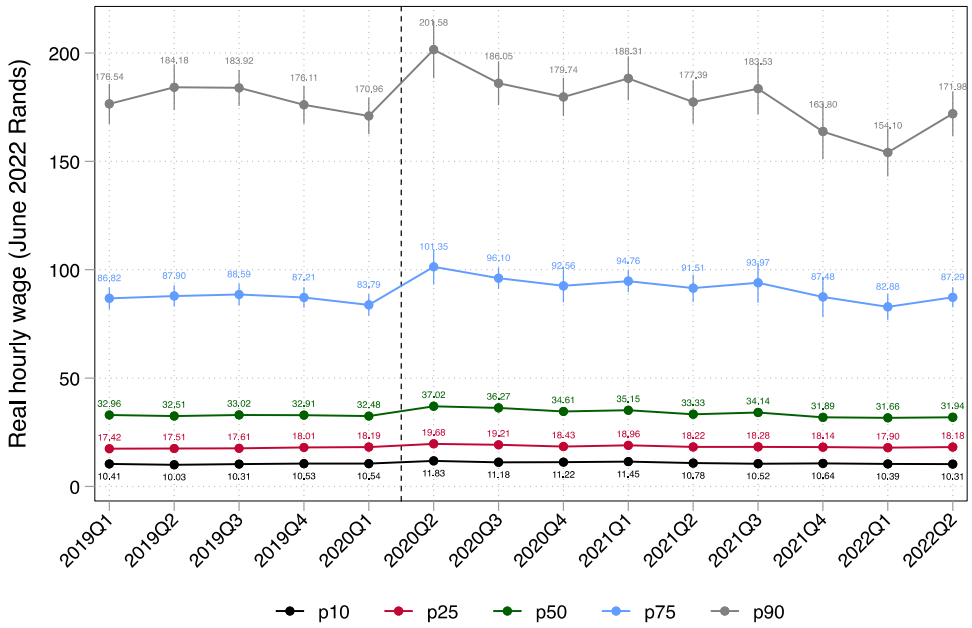


Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b; 2021b; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Wages adjusted for inflation and expressed in June 2022 Rands.

The observed rise in wages at the onset of the pandemic does not appear to be restricted to the mean but instead is observed across the entire wage distribution. Additionally, this change in wages appears to have been regressively distributed. Figure 9 presents the evolution of different percentiles of the wage distribution. First, the figure makes clear the extreme extent of wage inequality in the South African labour market even before the COVID-19 pandemic. Just prior to the pandemic, workers at the 10th percentile earned just R10.54 per hour, in contrast to the workers in the middle who earned more than 3 times more (R32.48 per hour). Inequality in the bottom half of the distribution is however far less severe than inequality in the top half, as documented in the literature. Workers at the 90th percentile

of the distribution earned R170.96 per hour – more than 5 times that of the median worker. At the onset of the pandemic, while wages increased at all percentiles considered, the change in wages was marginally higher towards the top of the distribution. At the top of the distribution, wages at the 90th percentile rose by 18 percent, in contrast to the middle where the median wage rose by 14 percent and the bottom where wages at the 10th percentile rose by 12 percent. All these differences are statistically significant by at least the 5 percent level. Thereafter, wages at all percentiles considered contracted towards their pre-pandemic levels and remained relatively stable for the remainder of the series.

Figure 9: Real hourly wage percentiles, 2019 – 2022

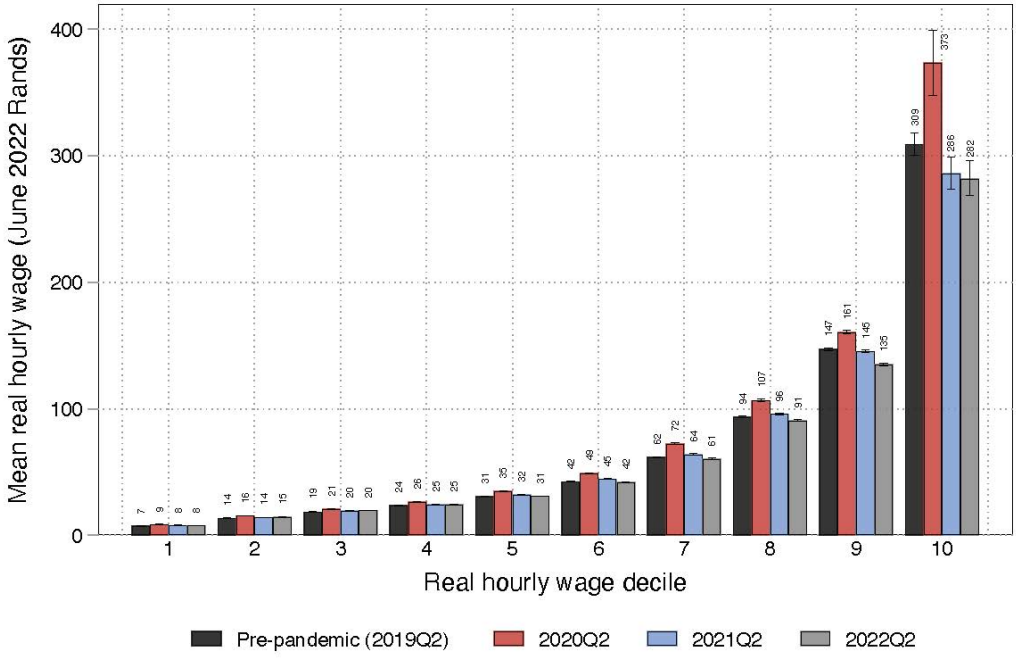


Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Spikes represent 95 percent confidence intervals.

Figure 10 plots estimates of decile-specific mean real hourly wages across the distribution and over time. The estimates again first describe the extreme extent of extreme wage inequality in the labour market, especially in the top half of the distribution. One year prior to the onset of the pandemic, the average worker among the poorest 10 percent of workers earned just R7 per hour, in contrast to the average worker in the middle who earned more than 4 times more (R31 per hour). As described above, inequality in the bottom half of the distribution is however far less severe than inequality in the top half. The average worker among the top decile of workers earned R309 per hour prior to the pandemic – nearly 10 and 44 times that of the average worker in the middle and bottom of the distribution. At the

onset of the pandemic in 2020Q2, mean wages in all deciles rose but to varying degrees. Relative increases ranged between 8 to 29 percent in the bottom half and 10 to 21 percent in the top half. All of these increases are statistically significant by at least the 5 percent level. During this quarter, however, the dispersion of the distribution was not considerably different. The ratio of mean wages at the middle compared to the bottom 10 percent remained at about 4, while that of the top 10 percent to the middle increased only marginally from 10 to 10.66. During the two years thereafter, these ratios remained relatively constant while most decile-specific mean wages reduced back to their pre-pandemic levels and were not statistically significantly different from them, however among the top 30 percent of workers, mean wages reduced further in real terms marginally below their pre-pandemic levels. Overall, these dynamics are consistent with the rightwards but transient shift in the distribution observed above and are suggestive of little to no change in wage inequality among the employed during the period.

Figure 10: Mean real hourly wages across the wage distribution, 2019 – 2022



Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b; 2021b; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Capped spikes represent 95 percent confidence intervals.

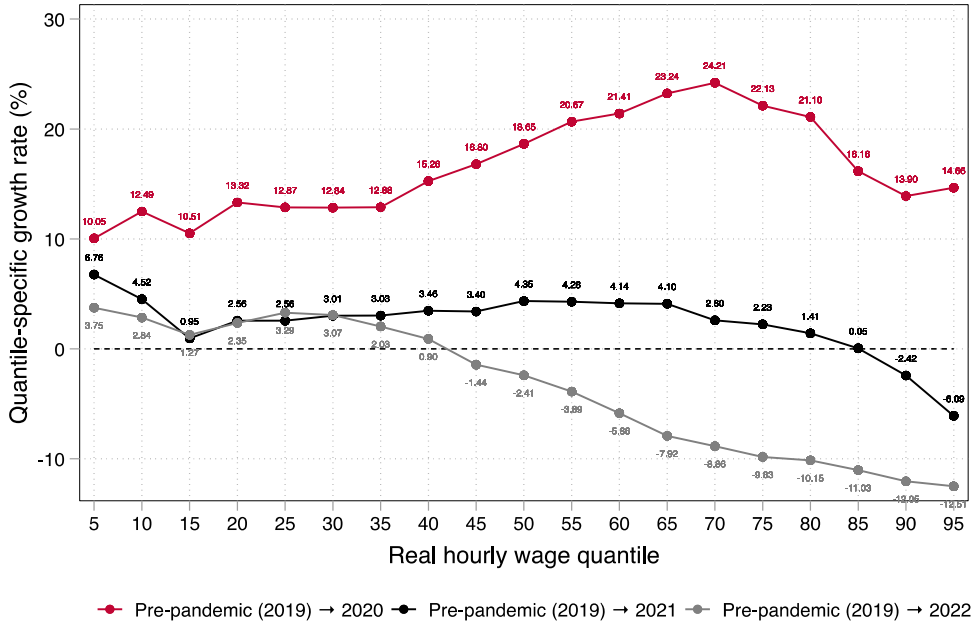
In brief, these estimates also point to a relatively large amount of minimum wage non-compliance in the labour market. Adjusted for inflation and using the National Minimum Wage (NMW) and relevant sectoral minimum wages for agriculture workers and domestic workers in place in January 2020, we estimate that just under one third (32.1 percent; s.e. = 0.6 percent) of employees earned sub-minimum

wages just prior to the pandemic in 2020Q1.[27] A large amount of non-compliance has also been reported by Bhorat et al. (2021) who however estimate a notably higher rate of 43.5 percent for the last quarter of 2019. This latter estimate is likely biased due to the use of the public QLFS wage data which includes StatsSA's imputations discussed in Section 3. This discrepancy in estimates is consistent with Kerr's (2022) analysis in an unpublished presentation which showed that the public QLFS data significantly overestimates minimum wage non-compliance. A more detailed analysis of minimum wage compliance during the pandemic in South Africa using the unimputed wage data here is beyond the scope of this paper, but certainly serves as an important area for future research.

The marginally regressive distribution of changes in real wages at the onset of the pandemic is again observed through the use of growth incidence curves – that is, a visual representation of quantile-specific growth rates across the wage distribution. We plot these curves in Figure 11 for three distinct periods to compare the evolution of unequal wage changes as the pandemic progressed. Considering the pre-pandemic period to the onset of the pandemic, all estimated growth rates are positive and exceed 10 percent after accounting for inflation, which is consistent with the previously observed rightwards shift in the distribution. During this period, growth rates were relatively constant up to the 35th quantile and thereafter rise until and inclusive of the 70th quantile. Growth rates reduce beyond this point but remain higher than those observed towards the bottom of the distribution. Real wages across most of the distribution were only marginally higher in 2021 one year after the pandemic's onset relative to the pre-pandemic period. One year thereafter, wages for approximately the bottom half of the distribution remained elevated but only marginally so, while those for the top half were lower. It should be noted that, to some extent, this contraction can be explained by the relative high consumer price inflation rates experienced during 2022 (Statistics South Africa, 2022c). Together, these dynamics reflect the temporary rightwards shift in the distribution at the pandemic's onset, followed by a relatively quick return to the pre-pandemic position thereafter.

[27] The NMW came into effect in January 2019, was set at R20 per hour excluding any allowances, bonuses, tips, or in-kind payments. It was applied across all sectors with the exceptions of agriculture workers, domestic workers, and public works programmes workers who were then entitled to minimum wages of R18, R15, and R11 per hour, respectively. Employers are also permitted to apply for exemptions in certain cases. Our calculation here accounts for both the NMW and sectoral minimum wages for agriculture workers and domestic workers. It however neither accounts for the public works minimum wage nor workers whose employers successfully applied for exemption. As such, the estimate may be biased upwards to some degree.

Figure 11: Growth incidence curves of real hourly wages, 2019 – 2022

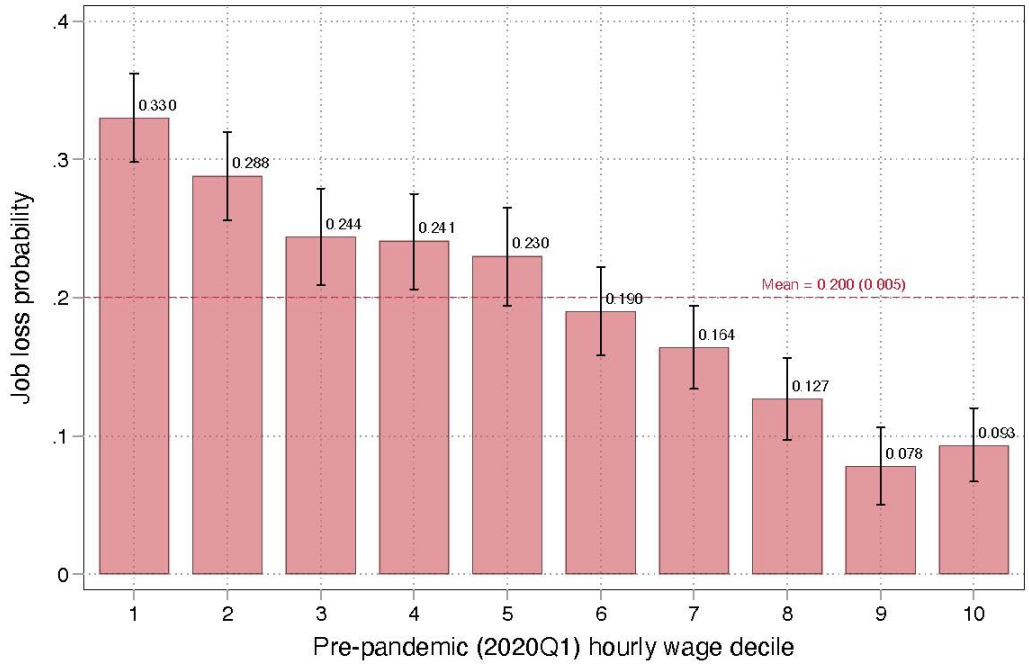


Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b; 2021b; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Wages adjusted for inflation and expressed in June 2022 Rands.

The increase in real wages at the pandemic's onset begs the questions of whether this was driven by workers receiving pay raises or alternatively reflects a compositional change in the employed population – that is, selection into remaining employed or conversely experiencing job loss across the wage distribution. We explore this mechanism by exploiting the unique panel nature of the QLFS data from 2020Q1 to 2020Q2, discussed in detail in the preceding paper, and estimating job loss probabilities (defined as being either unemployed, discouraged, or economically inactive in 2020Q2 conditional on being employed in 2020Q1) for the balanced panel sample across the pre-pandemic wage distribution. We present these estimates for each decile in Figure 12. First, it is apparent that workers across the entire distribution experienced job loss over this period. Second, job loss probabilities were notably heterogenous and regressive across the distribution. While the average worker faced a 20 percent chance of job loss, the steep and negative gradient with respect to pre-pandemic wages in the figure highlights the much greater vulnerability among lower-wage workers. A third (33 percent) of workers at the bottom of the distribution lost their jobs, in contrast to 23 percent of workers in the middle and 9 percent of workers at the top. The differences in these estimates are all statistically significant by at least the 5 percent level. This suggests that the observed increase in real wages did not occur because of pay raises but instead was due to regressive selection into remaining employed; hence, a 'composition' effect. In other words, higher-earning workers were more likely to remain employed relative to lower-

earning workers who dropped out of the wage distribution. This finding appears to have some external validity. Using an alternative panel dataset, Ranchhod and Daniels (2021) reached the same conclusion examining transitions out of employment from February 2020 (pre-pandemic) to April 2020 (the first month of the pandemic and associated lockdown in South Africa). South Africa is not unique in this regard, given that this mechanism has been shown to explain a rise in wages at the pandemic's onset in several other countries globally, including the United States (Cajner et al., 2020; Autor et al., 2023) and United Kingdom (Cribb et al., 2021).

Figure 12: Job loss probabilities by pre-pandemic real hourly wage decile, 2020Q1 – 2020Q2

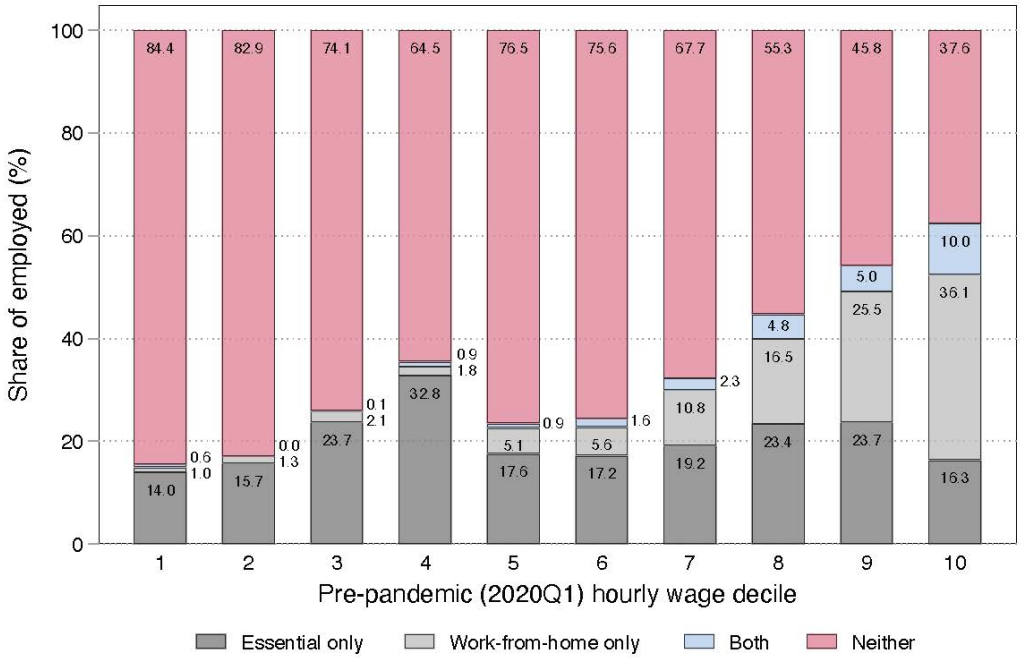


Author's own calculations. Source: QLFS 2020Q1, 2020Q2 (Statistics South Africa, 2020a; 2020b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) in the balanced panel sample who were employed in 2020Q1 but any labour market status in 2020Q2. Estimates weighted using sampling weights. Standard errors are adjusted for the complex survey design. Capped spikes represent 95 percent confidence intervals.

What might explain the higher job incidence among lower-wage workers observed above? One dominant mechanism the literature proposes is the distribution of workers who can and cannot continue to work given government-imposed, pandemic-related restrictions on economic activity; specifically, the distribution of workers in 'essential' jobs and those whose jobs allow them to work-from-home (WFH) (Baker, 2020; Dingel and Neiman, 2020; Kerr and Thornton, 2020; Mongey and Weinberg, 2020; Martin et al., 2022; Montenovo et al., 2022). While occupation and industry codes can be used to identify 'essential' workers, unfortunately prior to the pandemic no household survey in South Africa contained items related to a given worker's ability to WFH. As such, we follow the approach adopted by Kerr and Thornton (2020) who use disaggregated industry and occupation codes in the QLFS

to, first, identify workers in 'essential' jobs by cross-referencing the relevant government legislation,[28] and second, those who can plausibly WFH by following Dingel and Neiman (2020) and classifying occupations based on occupational context and activities using data from the Occupational Information Network (O*NET) dataset.[29] The interested reader is referred to Kerr and Thornton (2020) for a more detailed discussion of their approach. Figure 13 presents the relevant estimates across the pre-pandemic wage distribution in 2020Q1.[30] First, it is clear that lower-wage workers were significantly less likely than their high-earning counterparts to work in either work in 'essential' jobs or be able to WFH. Over 84 percent of the poorest decile of workers neither worked in 'essential' jobs nor could WFH, compared to 38 percent of the richest decile of workers. The probability of being able to WFH is also significantly higher among higher earners, likely because of the nature of tasks undertaken in these jobs. Overall, these estimates support the notion that the regressive distribution of job loss can be, at least in part, explained by lower-wage workers being less likely to work in 'essential' jobs or WFH.

Figure 13: Essential worker and work-from-home status by pre-pandemic real hourly wage decile



Author's own calculations. Source: QLFS 2020Q1 (Statistics South Africa, 2020a).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years). Estimates weighted using sampling weights. Categorisation of jobs into 'essential' and 'work-from-home' categories follows the approach employed by Kerr and Thornton (2020).

[28] Specifically, Government Gazette Numbers 11 062 and 11089.
[29] The O*NET is an occupational survey conducted by the U.S. Bureau of Labour Statistics. The authors' approach thus assumes equivalence in a given job's ability to be done from home in the US versus in South Africa. Because this need not be the case for certain jobs, such as teaching, the authors adjust the classification based on their own judgement of the South African context.
[30] Kerr and Thornton (2020) also examine the distribution of workers by 'essential' and work-from-home status across the wage distribution. However, the advantage of adopting their method here is that they could only make use of the inaccurate public QLFS wage data, whereas the analysis here uses the raw or observed QLFS wage data with multiple imputations.

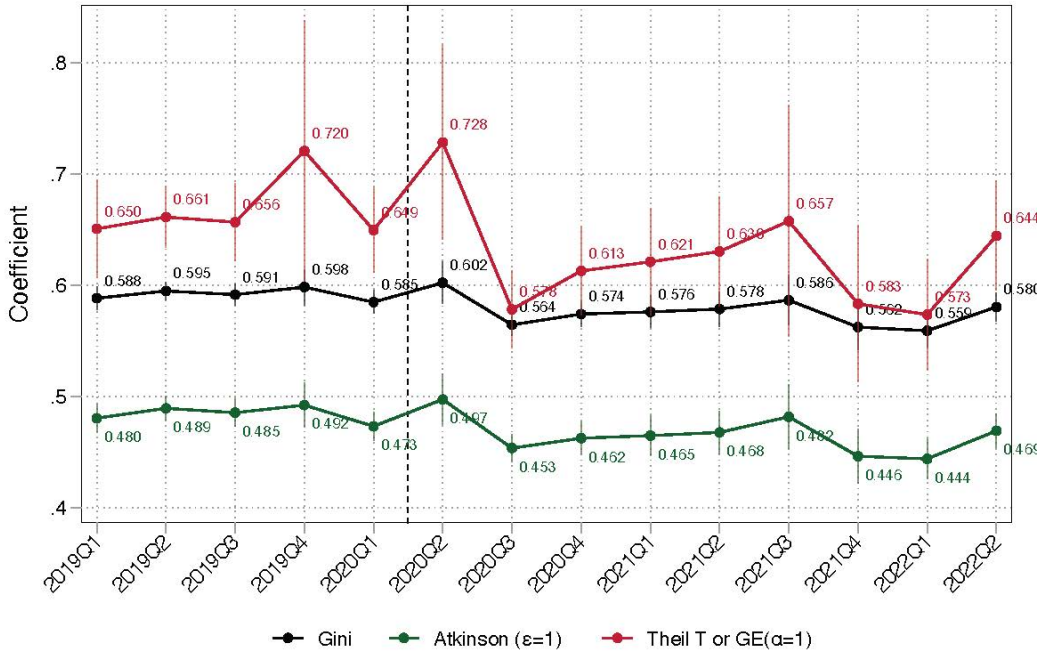
Together, the above estimates suggest that the regressive distribution of job loss mechanically drove the significant but transient rise in real wages across the distribution at the onset of the pandemic. However, despite this rightwards shift, the estimates so far do not point to a significant change in the dispersion of wages. To confirm these inequality dynamics, we now turn to estimating the aforementioned inequality indices across the series. Figure 13 presents the evolution of the estimated Gini coefficient, Atkinson index, and Theil T index. The estimates in the figure make is clear that, as described above, wage inequality prior to the pandemic was extremely high regardless of the measure, with estimated Gini, Atkinson, and Theil T coefficients of 0.585, 0.473, and 0.649 in 2020Q1, respectively. In the year prior to the pandemic, inequality remained relatively constant and only experienced marginal fluctuations, which is not necessarily surprising given that inequality indices are generally very slow-moving statistics (Cornia, 2014; Finn and Leibbrandt, 2018; Furceri et al., 2022).[31] The Theil T index serves as an exception given the large spike exhibited in 2019Q4. However, this spike appears to be driven by the inclusion of one observation with a particularly large, self-reported wage value.[32] While this wage was not detected as an outlier by the model described in Section 3, it's influence is notable. When this observation is excluded from the sample, the Gini and Atkinson indices remain relatively constant at 0.591 and 0.483, respectively, while the Theil T index reduces considerably to 0.666 – a level similar to the immediate preceding and proceeding survey waves. This latter estimate is also much more precisely estimated, with a confidence interval of a magnitude similar to neighbouring waves. Such outlying values are not evident in any other wave during the period, including at the pandemic's onset in 2020Q2 when the Theil T exhibits another increase.[33] This suggests that wage inequality prior to the pandemic was both relatively high and stable in the year preceding the pandemic.

At the pandemic's onset, the values of all indices rose but to varying degrees.[34] The Gini and Atkinson indices rose only marginally by 3 and 5 percent, respectively. These increases are however only marginally larger than those observed one year prior. On the other hand, the Theil T index experienced a larger jump of 12 percent – seven times larger than the increase during the equivalent period one year prior. Considering the sensitivity of this measure to wage changes towards the top of the distribution, these dynamics are consistent with the prior observation of larger wage changes towards the top. Thereafter, all indices indicate that wage inequality reduced to below pre-pandemic levels in

[31] A temporary spike of the Theil T estimate in 2019Q4 serves as the exception. Such a dynamic is not however shared by the other inequality indices. The estimate is also much less precisely estimated, as indicated by the wide confidence intervals, and is not statistically significantly different from the immediate preceding or proceeding estimates.
[32] The referenced worker reported an hourly wage of approximately R5 792 in real terms, which significantly exceeds the maximum self-reported wage for both the preceding and proceeding periods (R2 134 and R4 664, respectively).
[33] The maximum self-reported wage in 2020Q2 is R2 604.
[34] Recall that higher values for all indices indicate greater inequality.

the following quarter (2020Q3) before gradually rosing at similar rates thereafter. This points to the transient nature of the rise in wage inequality at the pandemic's onset. Inequality experienced another reduction from 2021Q3 to 2022Q1 before rising to again to the pre-pandemic level again by the end of the period. It should be noted that these levels and trends are very insensitive to our treatment of furloughed workers, as shown in Figure A1 in the appendix which presents the equivalent estimates when these workers are excluded from the sample.

Figure 14: Relative wage inequality estimates by measure, 2019Q1 – 2022Q2

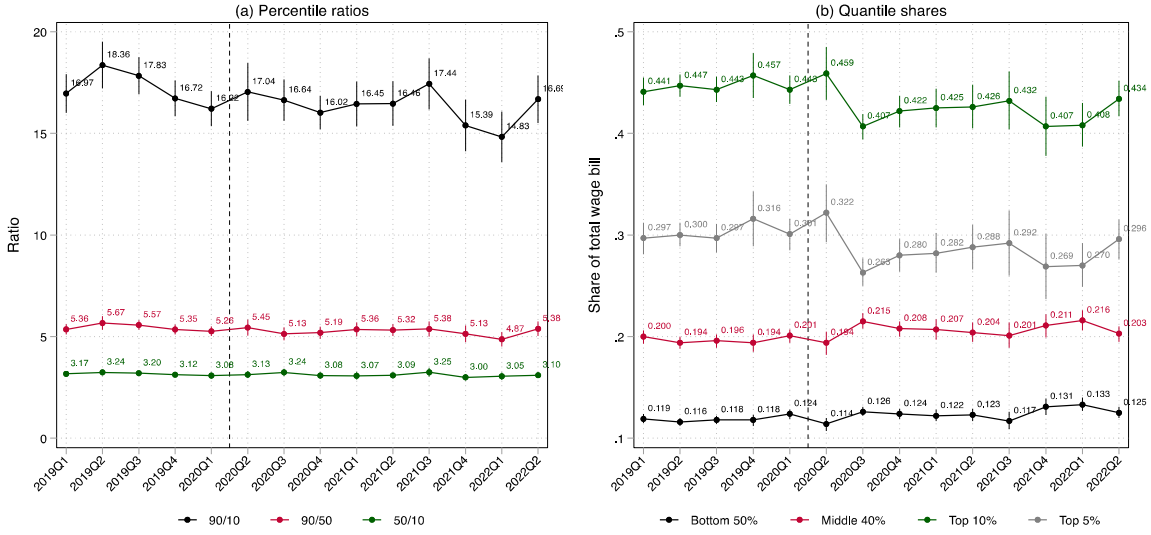


Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Spikes represent 95 percent confidence intervals.

we now focus on specific parts of the distribution by estimating and analysing the evolution of percentile ratios and quantile shares. These estimates are presented in Figure 14. First, panel (a) again highlights the greater amount of inequality in the top half of the distribution relative to the bottom half. Just before the pandemic, workers at the 90th percentile earned more than 16 times that of workers at the bottom (10th percentile) of the distribution and more than 5 times that of workers in the middle. These latter workers earned just over 3 times that of workers at the 10th percentile, highlighting the relative compression of wages towards the bottom. Notably, these estimates suggest that wage inequality was gradually reducing during the year preceding the pandemic, particularly inequality between the bottom and top of the distribution. From 2019Q2 to 2020Q1, the 90/10 ratio contracted by 13 percent from 18.4 to 16, while the 90/50 ratio also reduced but at a nearly 50 percent slower rate. Panel (b) tells a

similar story of extreme and persistent wage inequality but from the perspective of income concentration. Prior to the pandemic, the top 10 percent of workers accounted for 44 percent of all wages earned in the labour market, while the bottom 50 percent accounted for just 12 percent. Within the top 10 percent, wages were concentrated among the top 5 percent who accounted for 30 percent of all wages, or over two-thirds percent of all wages within the top decile.

Figure 15: Wage percentile ratios and quantile shares, 2019Q1 – 2022Q2



Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Spikes represent 95 percent confidence intervals.

At the pandemic's onset, the gap between workers at the top and bottom widened marginally, with 90th percentile workers now earning 17 times that of 10th percentile workers. However, this difference is not statistically significant. Statistically insignificant changes are also observed for the 90/50 and 50/10 ratio. It is unsurprising then that the quantile shares of workers towards the top of the distribution grew, albeit insignificantly, while those towards the bottom shrunk.[35] As the pandemic progressed into the next quarter (2020Q3), wage inequality reduced to below pre-pandemic levels as previously observed in Figure 13. This transient contraction is particularly evident when considering the quantile shares as opposed to percentile ratios. From 2020Q2 to 2020Q3, the top decile's share reduced by 46 to 41 percent while, concurrently, the bottom 50 percent's share grew from 11 to 13 percent and the middle 40 percent's from 19 to 22 percent. Thereafter, wage inequality gradually returned to levels similar to

[35] This latter contraction in the bottom 50 percent's share of one percentage point is statistically significant at the 5 percent level.

the pre-pandemic period, as observed in Figure 13 and again highlighting the transient nature of the rise in wage inequality. As such, these estimates make it clear that extreme wage inequality persisted even as the labour market was recovering with respect to job loss.

Importantly, because the inequality estimates presented in Figures 13 and 14 are based on cross-sectional samples of the employed, they do not explicitly account for selection into remaining employed at the pandemic's onset; in other words, a composition effect brought about by an abrupt and regressive distribution of job loss which resulted in an over two million workers – as shown in the preceding paper – being effectively removed from the wage distribution. Accounting for this composition effect would entail retaining the previously employed in the sample and regarding them as zero wage earners. To do so, we adopt two approaches. First, we make use of a cross-sectional recall item in the survey – which asks the unemployed, conditional on having ever worked before, how long ago it was since they last worked – to identify those who were employed in 2020Q1 just prior to the pandemic but unemployed thereafter.[36] Observations beyond 2020Q4 are not considered because the available response items do not allow one to identify those previously employed just prior to the pandemic. Second, we exploit the pandemic-induced change to the survey design which resulted in it becoming an unbalanced panel survey from 2020Q1 to 2021Q1, as discussed in the preceding paper, and make use of household and person identifiers as well as observable covariates to identify those who were employed in 2020Q1 but unemployed thereafter.[37] While attrition results in the sample obtained from this approach not including all observations interviewed in 2020Q1, which may be cause for concern for bias, the estimates are very similar in both magnitude and precision to those obtained using the cross-sectional approach described above, as shown later. For both these approaches then, the sample in a given wave comprise the employed and those previously employed just prior to the pandemic, with the latter's wages being set to zero.

Figure 15 presents estimates of these 'composition-controlled' Gini coefficients using the above two approaches for 2020Q1 to 2020Q4. For comparison, these are plotted alongside estimates using the

[36] Possible responses to this item include "less than 3 months"; "3 months to less than 6 months"; "6 months to less than 9 months"; "9 months to less than 1 year"; "1 year to less than 3 years"; "3 years to 5 years"; "more than 5 years"; and "don't know". Here, unemployed observations in 2020Q2 were regarded as employed prior to the pandemic if they reported being employed either "less than 3 months" or "3 months to less than 6 months" ago, those in 2020Q3 were regarded as employed prior to the pandemic if they reported being employed either "3 months to less than 6 months" or "6 months to less than 9 months" ago, and those in 2020Q3 were regarded as employed prior to the pandemic if they reported being employed either "6 months to less than 9 months" or "9 months to less than 1 year" ago.

[37] Because of the anonymity of observations in the survey, covariates in addition to household and person identifiers were used to ensure the same individual was being observed over time. These included self-reported race, sex, and age. Age was permitted to vary by one year across a given quarter-by-quarter pair. This approach resulted in 19 943 unique observations observed four times from 2020Q1 to 2020Q4.

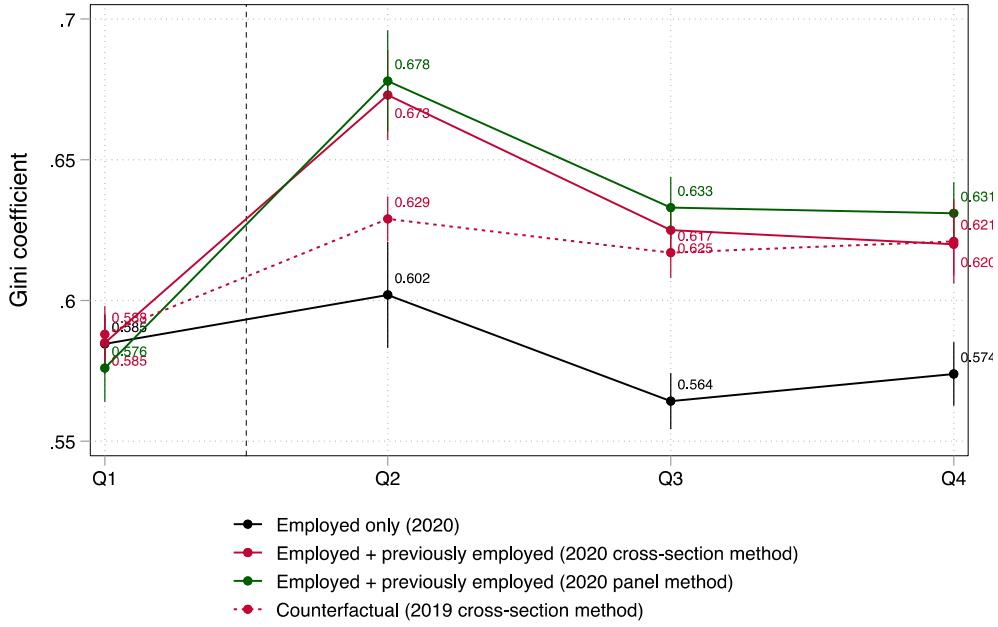
sample of the employed only, which is equivalent to the sample used previously in Figures 13 and 14. However, inferring that any wave-specific difference between a 'composition-controlled' coefficient and the 'employed only' coefficient is attributable to pandemic-induced job loss may be inaccurate because of other events that would have happened during the same period in the pandemic's absence. For instance, the transition from employment just prior to the pandemic to unemployment thereafter may simply be the consequence of seasonality effects. To account for this, we include estimates from a sample derived using the cross-sectional method described above but on 2019 data. Because this sample comprises a similar sample as that obtained using the cross-sectional method for 2020 but just for one year prior, we explicitly assume that any difference between the 'composition-controlled' estimates across 2019 and 2020 is attributable to the pandemic, and hence refer to an estimate obtained from it as the counterfactual of the 'composition-controlled' Gini for 2020, interpreted as what the 'composition-controlled' Gini may have been in the absence of the pandemic.

The estimates show that, after accounting for a change in the composition of workers, the pandemic increased wage inequality significantly at its onset. While the Gini coefficient using the cross-sectional employed samples increased from 0.585 in 2020Q1 by just 3 percent to 0.602 in 2020Q2, as mirrored in Figure 13, the 'composition-controlled' coefficients rose by five to six times faster depending on the method. Using the cross-sectional method, the coefficient grew by 15 percent to 0.673, compared to a rise of 18 percent to 0.678 when the panel method is alternatively used.[38] These latter estimates for 2020Q2 are not statistically significantly different from one another. This partially reflects a reduction in the median hourly wage by 17 percent to R26.95 when the previously employed are included, as opposed to rising by 14 percent to R37.02 when they are excluded as shown in Figure 9. On the other hand, using the cross-sectional method but on data from the same period one year prior, the counterfactual estimate also rose from a similar base (in terms of both statistical significance and magnitude) but by a more than 50 percent lower rate (7 percent) during the same period.[39] Thereafter, the 'composition-controlled' coefficients gradually reduced toward their pre-pandemic levels while the counterfactual coefficient remained relatively constant. In both 2020Q3 and 2020Q4, while the 'composition-controlled' coefficients using the cross-sectional method were statistically insignificantly different from the counterfactual estimates, those derived using the panel method were higher but only marginally so. This observation is consistent with the prior finding that higher wage inequality at the pandemic's onset appears to have only been transient.

[38] The 2020Q1 Gini coefficients using the 'employed only' and 'employed + previously employed (2020 cross-sectional method)' samples are identical because, by construction, both make use of the same sample of workers in the wave, while the coefficient using the 'employed + previously employed (2020 panel method)' sample is marginally but not statistically significantly lower because the balanced panel sample is used.

[39] This rate is of course more than twice the growth rate in the Gini when the employed cross-section samples are used.

Figure 16: Gini coefficient estimates accounting for a composition effect, by sample

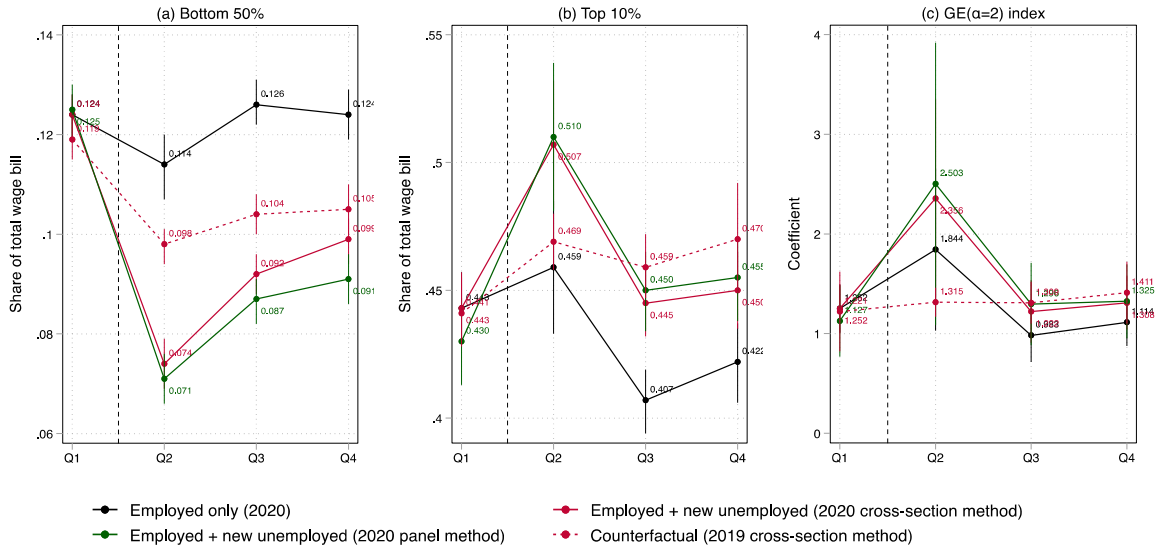


Author's own calculations. Source: QLFS 2019Q1 - 2020Q4 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) throughout but varies as follows: "Employed only (2020)" includes the employed for each cross-section; "Employed + previously employed (2020 cross-section method)" includes the employed for each cross-section as well as those previously employed during the pre-pandemic period (2020Q1) using the described cross-section method; "Employed + previously employed (2020 panel method" includes the employed for each cross-section as well as those previously employed during the pre-pandemic period (2020Q1) using the balanced panel data; "Counterfactual (2019 cross-section method)" includes the equivalent sample for "Employed + previously employed (2020 cross-section method)" but using the 2019 data. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Spikes represent 95 percent confidence intervals.

Assuming the 2019 'composition-controlled' Gini estimates do indeed serve as an appropriate counterfactual, the implications of these trends are four-fold. First, they show that not accounting for the change in the composition of workers may lead to misinterpretations of wage inequality dynamics during this period. Second, they suggest that wage inequality may have risen anyway in the pandemic's absence, but not to the same extent. Third, they suggest that approximately half of the observed rise in the 'composition-controlled' Gini at the pandemic's onset is explained by the pandemic itself, or in other words, the pandemic itself increased wage inequality by between 7 – 8 percent or 4.4 – 4.9 Gini points in the immediate term. Finally, this rise in wage inequality appears to have been temporary, with estimates from 2020Q3 onwards being only marginally different than what they may have been in the pandemic's absence. These dynamics appear largely insensitive to the chosen measure of inequality. The trajectory of the 'composition-controlled' and counterfactual Gini estimates presented in Figure 15 closely mirror the equivalent trends for top 10 and bottom 50 percent quantile shares as well as the General Entropy (GE) measure presented in Figure 16. Regarding the latter, recall that the Theil T index, equivalent to GE($\alpha = 1$), exhibited a larger jump than other indices at the pandemic's onset as shown

in Figure 13, implying larger wage changes towards the top of the distribution. While the Theil T cannot be estimated here,[40] the equivalent trends using $\alpha = 2$ shown in panel (c) reveal an even larger increase at the pandemic's onset relative to the $\alpha = 1$ case. This is not necessarily surprising given the prior observation of larger observed wage changes towards the top of the distribution, and the positive relationship between positive α values and the sensitivity of this measure to such changes.

Figure 17: Quantile share and general entropy coefficient estimates accounting for a composition effect, by sample



Author's own calculations. Source: QLFS 2019Q1 - 2020Q4 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) throughout but varies as follows: "Employed only (2020)" includes the employed for each cross-section; "Employed + previously employed (2020 cross-section method)" includes the employed for each cross-section as well as those previously employed during the pre-pandemic period (2020Q1) using the described cross-section method; "Employed + previously employed (2020 panel method" includes the employed for each cross-section as well as those previously employed during the pre-pandemic period (2020Q1) using the balanced panel data; "Counterfactual (2019 cross-section method)" includes the equivalent sample for "Employed + previously employed (2020 cross-section method)" but using the 2019 data. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Spikes represent 95 percent confidence intervals.

5.2. Decomposition analysis of temporal wages changes

5.2.1. At the mean: Oaxaca-Blinder estimates

In this final component of this paper's analysis, we present the results of our decomposition analysis of the structural and compositional drivers of the changes in wages and wage inequality from before to after the onset of the pandemic, both at the mean and across the entire wage distribution using OB and

[40] The Theil T index cannot be estimated because any GE measure with $\alpha < 2$ is undefined in the presence of non-positive wage values, which are explicitly included for the previously employed here.

RIF decomposition, respectively. We begin with the analysis at the mean and present the results from the overall and detailed OB decompositions in Tables 5, 6, and 7. Table 5 reports the mean real hourly wages (on a logarithmic scale) in a given first and second period, the temporal difference, and how much this difference is explained by composition (that is, changes in the distribution of covariates) and structure (that is, changes in the associated returns to these covariates) effects. Table 6 considers the detailed decomposition of this full composition effect into the contributions from each group of covariates, while Table 7 does the same but for the full structure effect.

The overall decomposition results presented in Table 5 make it clear that the increase in the mean wage from the pre-pandemic period to after the pandemic's onset was primarily driven by a composition effect. As shown in column (1) the log mean wage increased by 0.133 log points, which is expected for reasons discussed in the preceding section, and while both a composition and structural effect explains this change, most (71 percent) is explained by a composition effect. This latter finding is consistent with our previous finding that the rise in the mean wage over this period was driven by a compositional shift in the employed population; that is, lower-wage workers were significantly more likely to experience job loss and hence drop out of the wage distribution. Although the composition effect is dominant, what is also notable, however, is the non-negligible magnitude of the structure effect. Approximately 29 percent of the increase in the average wage is explained by changes in the associated returns to individual-level characteristics.

Overall, the results in columns (2) and (3) imply that as the pandemic progressed and the labour market partially recovered until 2022Q2, the employed population returned to a similar composition compared to the pre-pandemic period, however concurrently, the difference in associated returns to individual-level characteristics over the period grew. As shown in column (2) which compares wages one year after the pandemic's onset to those in the pre-pandemic period, the mean wage reduced to be marginally higher than its pre-pandemic level but the two estimates are not statistically significantly different from each other. The composition effect, although less than half the magnitude of the effect in the preceding period, was also dominant during this period, reflecting another compositional change in the employed population as the labour market recovered. On the other hand, the magnitude of the structure effect was relatively constant compared to the preceding period but changed in sign, and hence partially offset the positive compositional effect. Another year later, as shown in column (3), the mean wage marginally reduced further and the difference compared to the pre-pandemic level remained insignificant. The composition effect reduced further in magnitude and became only marginally significant, while the structure effect estimate grew by 30 percent to-0.044 and remained highly significant.

Table 5: Overall Oaxaca-Blinder decomposition estimates of changes in mean real hourly wages, by period

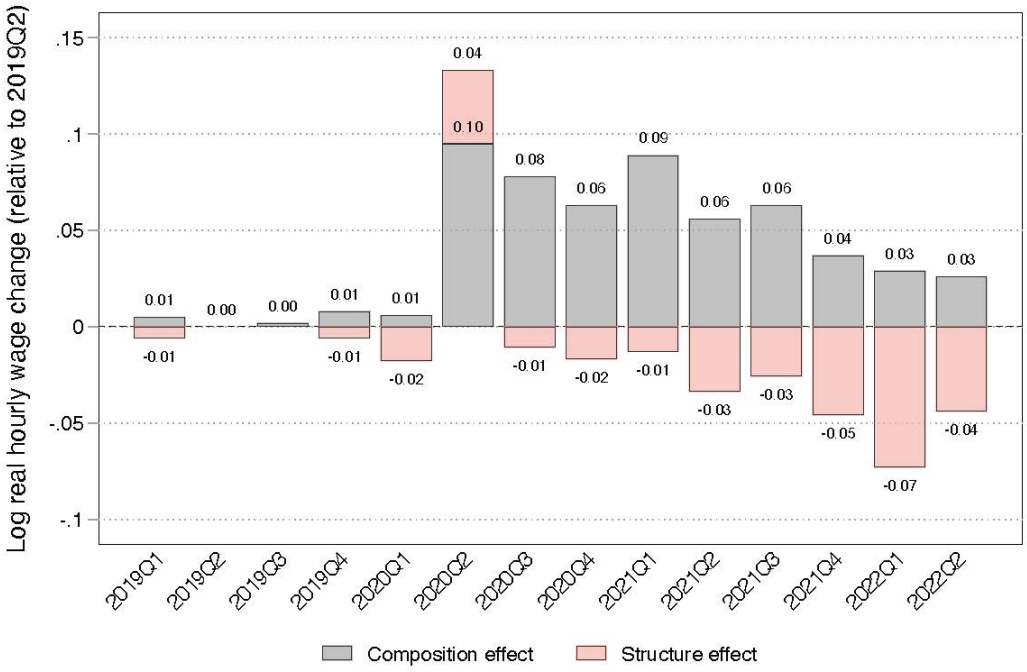
	(1) Pre-pandemic (2019Q2)-2020Q2	(2) Pre-pandemic (2019Q2)-2021Q2	(3) Pre-pandemic (2019Q2)- 2022Q2
Pre mean (real hourly wage, log scale)	3.645***	3.645***	3.645***
	(0.014)	(0.015)	(0.015)
Post mean (real hourly wage, log scale)	3.778***	3.667***	3.628***
	(0.019)	(0.018)	(0.017)
Difference	0.133***	0.022	-0.017
	(0.020)	(0.021)	(0.020)
Composition	0.095***	0.056***	0.026*
	(0.015)	(0.016)	(0.016)
Structure	0.038***	-0.034**	-0.044***
	(0.014)	(0.014)	(0.013)
Observations	26 735	28 572	29 644

Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b;2021b; 2022b).

*Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design and are presented in parentheses. Hourly wages adjusted for inflation and expressed in June 2022 Rands. * p<0.10; ** p<0.050; *** p<0.010.*

The above finding is insensitive to the specific wave-to-wave pairs selected above. Figure 17 presents estimates of the full composition and structure effects for the entire period. Each temporal change is relative to the same baseline period – 2019Q2 – to be consistent with the estimates above. The estimates show that the real mean wage was relatively constant prior to the pandemic, with both composition and structure effect estimates being statistically insignificant and close to zero in magnitude. As shown above, at the pandemic's onset both a composition and structure effect drove the rise in the mean wage, however the former effect was dominant. During the two years thereafter the magnitude of the composition effect reduced in size, reflecting a growing similarity of the characteristics of workers compared to the pre-pandemic period as the labour market recovered and jobs were re-gained. Concurrently, however, the size of the structure effect gradually grew and was larger than the only marginally significant composition effect by the end of the period, indicative of a significant change in the associated wage returns to individual-level characteristics.

Figure 18: Overall Oaxaca-Blinder decomposition estimates of changes in mean real hourly wages over the whole period



Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights and are adjusted for the complex survey design. Hourly wages adjusted for inflation and expressed in June 2022 Rands.

we now consider the detailed decomposition of this full composition effect into the contributions from each group of covariates, as shown in Table 6. As shown in column (1), the estimates show that at the pandemic's onset, five specific covariates – trade union membership, main occupation, years of education, formal sector employment, and public sector employment, in order of magnitude – significantly explain the full composition effect. Together, these explain about 95 percent of the full composition effect, and hence over two-thirds (68 percent) of the rise in the real mean wage. It is notable that no demographic variables explain the composition effect at the mean. The coefficients on all covariate groups are positive and highly significant, indicating that after the pandemic's onset the composition of workers were more unionised, in typically higher-paying occupations, more educated, and more likely to work in the formal and public sectors. These shifts are consistent with the compositional shifts observed in the preceding paper and, together with the significant amount of job loss observed during the period, imply that workers with these characteristics were simply more likely to remain employed. During the two years thereafter as the real mean wage mechanically returned to the pre-pandemic level as employment recovered and the magnitude of the full composition effect approached zero, the significance of all but one of these covariates disappeared. The coefficient on

education remained significant and of the same sign and similar magnitude to the preceding periods, reflecting a marginally more educated worker population.[41]

Table 6: Detailed Oaxaca-Blinder decomposition estimates of composition effect, by period

	(1) Pre-pandemic (2019Q2)-2020Q2	(2) Pre-pandemic (2019Q2)-2021Q2	(3) Pre-pandemic (2019Q2)- 2022Q2
Race	0.003	0.000	-0.003
	(0.005)	(0.005)	(0.005)
Age	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
Province	0.001	0.001	-0.001
	(0.002)	(0.002)	(0.002)
Female	0.000	0.000	-0.002
	(0.001)	(0.001)	(0.001)
Urban	-0.001	-0.001	0.000
	(0.001)	(0.001)	(0.001)
Education	0.024***	0.017***	0.027***
	(0.004)	(0.004)	(0.004)
Public sector	0.005***	0.000	0.002
	(0.002)	(0.001)	(0.001)
Formal sector	0.010***	0.000	-0.001
	(0.002)	(0.003)	(0.002)
Experience	0.002	0.005***	-0.002
	(0.002)	(0.002)	(0.002)
Unionisation	0.031***	0.022***	0.001
	(0.004)	(0.004)	(0.003)
Industry	0.001	0.003	-0.001
	(0.002)	(0.003)	(0.002)
Occupation	0.018***	0.007	0.004
	(0.007)	(0.007)	(0.006)
Observations	26 735	28 572	29 644

Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b;2021b; 2022b).

*Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design and are presented in parentheses. Hourly wages adjusted for inflation and expressed in June 2022 Rands. Decomposition for categorical variables (industry, occupation, race, age, and province) based on "normalized" effects; that is, effects are expressed as deviation contrasts from the grand mean. Reference groups for categorical variables as follows: Province: Western Cape; Age: 15-34 years; Race: African/Black; Occupation: Managers; Industry: Agriculture, forestry, and fishing. * p<0.10; ** p<0.050; *** p<0.010.*

Similarly, Table 7 presents the detailed decomposition of this full structure effect. The estimates in column (1) show that just one covariate had a statistically significant and positive coefficient at the onset of the pandemic: main industry of employment. This is indicative of a change in sectoral wage premia during the period, at least at the mean, and may simply reflect varying sectoral returns for job-retainers relative to job-losers, or alternatively variation in which sectors were legally permitted to operate during

[41] Mean years of education was 11.39 years in 2022Q2 compared to 11.06 years in 2019Q2, a marginal but statistically significant difference at the 1 percent level.

the hard lockdown period at the pandemic's onset. As with the composition effect, it is notable that no demographic variables explain the structure effect at the mean. As shown in column (2), such varying returns however disappear one year later, and instead only the coefficient on the urban indicator is significant. One further year later, as shown in column (3), the significance of this estimate also disappears, leaving all estimates insignificant. Despite this, the full structure effect estimate in Table 5 is highly significant, implying differences in the associated returns to various characteristics in 2022Q2 relative to the pre-pandemic period. Solely considering coefficient magnitudes suggests these might be related to education and potential experience, however the inflated standard errors do not allow me to arrive at such a conclusion confidently. As such, while differences in associated returns appear to exist between the two periods, the data does not enable one to identify the covariates these returns pertain to.

Table 7: Detailed Oaxaca-Blinder decomposition estimates of structure effect, by period

	(1) Pre-pandemic (2019Q2)-2020Q2	(2) Pre-pandemic (2019Q2)-2021Q2	(3) Pre-pandemic (2019Q2)- 2022Q2
Race	0.000	-0.006	0.012
	(0.022)	(0.027)	(0.033)
Age	0.005	0.033	-0.003
	(0.028)	(0.026)	(0.028)
Province	-0.005	-0.006	-0.012
	(0.010)	(0.009)	(0.010)
Female	-0.012	-0.011	-0.012
	(0.012)	(0.014)	(0.012)
Urban	-0.042	-0.051**	-0.033
	(0.026)	(0.023)	(0.025)
Education	0.042	0.047	0.049
	(0.075)	(0.074)	(0.066)
Public sector	0.000	-0.011	-0.014
	(0.008)	(0.007)	(0.009)
Formal sector	-0.047	-0.008	-0.037
	(0.035)	(0.036)	(0.032)
Experience	-0.023	-0.037	-0.063
	(0.070)	(0.074)	(0.074)
Unionisation	-0.015	0.009	0.003
	(0.012)	(0.013)	(0.013)
Industry	0.038**	0.000	-0.003
	(0.016)	(0.014)	(0.014)
Occupation	-0.010	-0.013	-0.010
	(0.023)	(0.019)	(0.018)
Observations	26 735	28 572	29 644

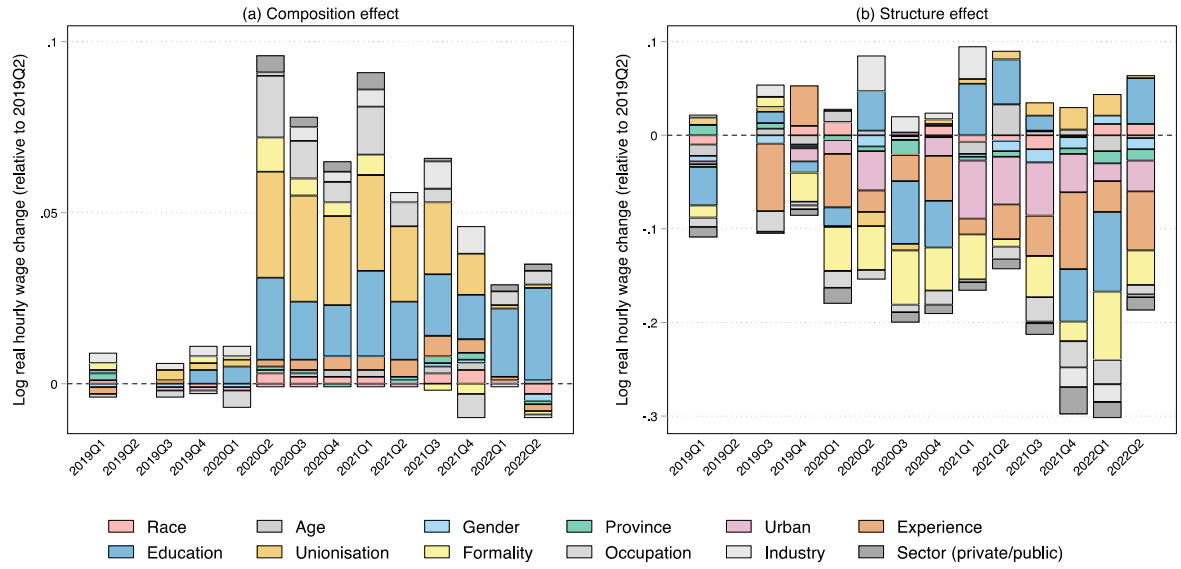
Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b;2021b; 2022b).

*Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design and are presented in parentheses. Hourly wages adjusted for inflation and expressed in June 2022 Rands. Decomposition for categorical variables (industry, occupation, race, age, and province) based on "normalized" effects; that is, effects are expressed as deviation contrasts from the grand mean. Reference groups for categorical variables as follows: Province: Western Cape; Age: 15-34 years; Race: African/Black; Occupation: Managers; Industry: Agriculture, forestry, and fishing. Constant term omitted. * p<0.10; ** p<0.050; *** p<0.010.*

These detailed decomposition estimates also appear to be insensitive to the specific wave-to-wave pairs selected above. Similar to Figure 17, Figure 18 presents estimates of the detailed composition and structure effects for the entire period, with each temporal change again being relative to the same baseline period (2019Q2). The estimates in panel (a) again show how trade union membership, main occupation, education, formal sector employment, and public sector employment primarily explain the full composition effect, both at the pandemic's onset and beyond. The influence of all other covariates were both economically and statistically insignificant. The influence of most of these covariates reduced meaningfully or fell away completely by the end of 2021, with the exception of education which served as the only covariate which persisted in influence throughout the period, again reflecting a marginally more educated worker population. The narrative pertaining to the detailed structure effects is less clear. As shown in panel (b), this is primarily because the majority of estimates are very small in magnitude and vary in sign.[42] The implication of this is that overall, as mentioned above, while differences in associated returns appear to exist particularly between the last period and the pre-pandemic period, the data does not enable one to identify the covariates these returns are with respect to.

[42] It should be noted that each wave-specific constant term of the OB decomposition of the full structure effect is omitted from the figure here. In all waves the constant term is positive and relatively large, and hence the full structure effect is relatively small when summing the constant with covariate groups coefficients.

Figure 19: Oaxaca-Blinder detailed decomposition of composition and structure effects for the whole period



Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design and are presented in parentheses. Hourly wages adjusted for inflation and expressed in June 2022 Rands. Decomposition for categorical variables (industry, occupation, race, age, and province) based on "normalized" effects; that is, effects are expressed as deviation contrasts from the grand mean. Reference groups for categorical variables as follows: Province: Western Cape; Age: 15-34 years; Race: African/Black; Occupation: Managers; Industry: Agriculture, forestry, and fishing. Constant term omitted from structure effect decomposition estimates.

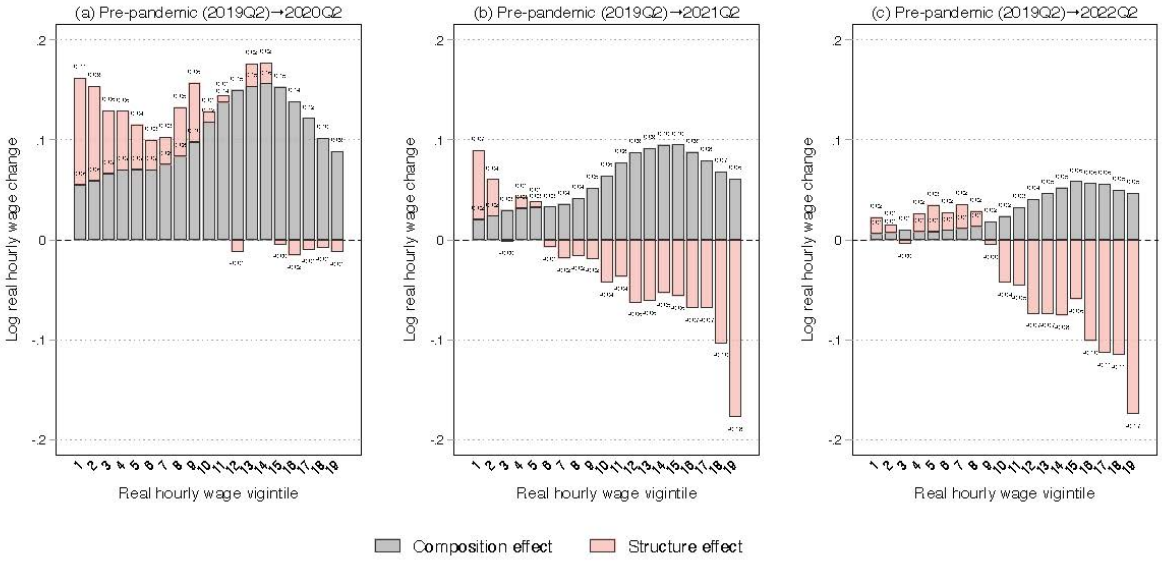
5.2.2. Across the distribution: Recentered Influence Function estimates

we now move beyond the mean and examine the results of the RIF decomposition of wage changes from before to after the onset of the pandemic across the entire wage distribution. Figure 19 presents a visual representation of the estimates of the change in real hourly wages, again on a logarithmic scale, and the decomposed contributions of the full composition and structure effects for each quantile of the wage distribution, again for three distinct periods. The estimates are presented in the form of a stacked bar chart for ease of interpretation; that is, the net change in the log wage for a given quantile is equivalent to the sum of the individual components.

The estimates reveal significant heterogeneity in both the magnitude and direction of the full composition and structure effects across the wage distribution, highlighting the advantage of RIF decomposition over OB decomposition. Overall, at the pandemic's onset, a change in the characteristics of workers primarily explain the observed rise in real wages across most of the distribution – which is in line with the analysis at the mean – apart from, however, at the very bottom where a change in the

returns to these characteristics primarily explain this change. As shown in panel (a), real wages were higher at the pandemic's onset relative to one year prior across the entire distribution, which is consistent with the growth incidence curve estimates in Figure 11, reflecting the regressive distribution of job loss during the period. In the average vigintile, the composition effect explains 80 percent of the rise in wages, from 52 percent in the 3rd vigintile at the bottom to 92 percent in the middle and exceeding 100 percent towards the top. As in the mean case, both the composition and structure effects are positive across most of the distribution.[43] The structure effect however plays a relatively negligible role but appears to grow in magnitude with lower wages and actually dominates the composition effect in the lowest decile. This suggests again that, whereas changes in the composition of workers primarily explains wage increases across most of the distribution, changes in the returns to these characteristics explain wage increases at the very bottom.

Figure 20: Recentered Influence Function decomposition of total wage change into composition and structure effects across the wage distribution, by period



Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b;2021b; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Hourly wages adjusted for inflation and expressed in June 2022 Rands. Decomposition for categorical variables (industry, occupation, race, age, and province) based on "normalized" effects; that is, effects are expressed as deviation contrasts from the grand mean.

Panel (b) shows that, as the pandemic progressed and the labour market partially recovered one year following the pandemic's onset, the magnitude of the composition effect reduced, reflecting fewer differences in the profile of workers relative to the pre-pandemic period. However, the composition

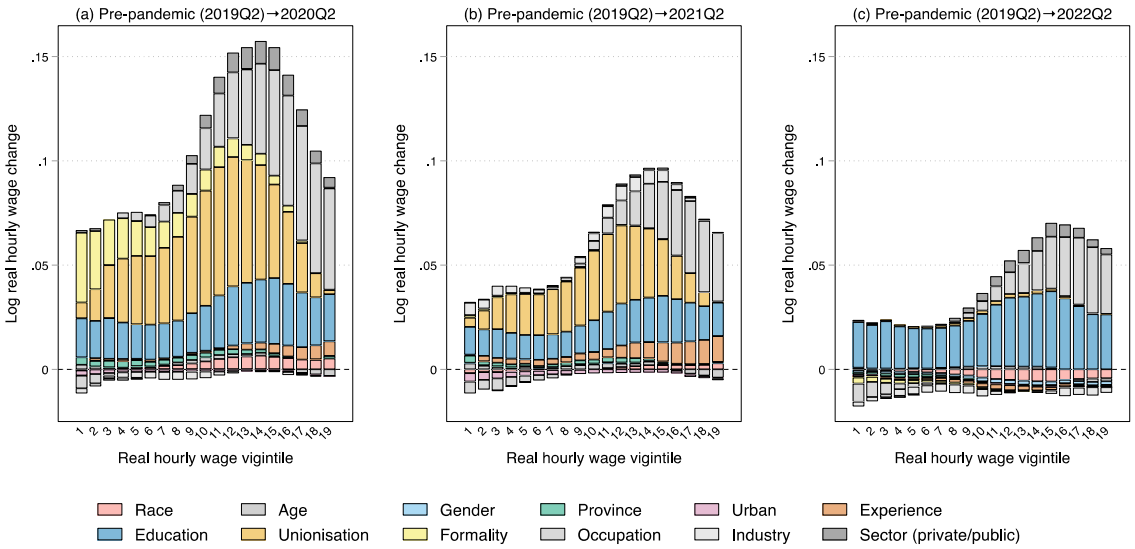
[43] The structure effect is negative at top of the wage distribution, partially offsetting the composition effect, however the estimates are negligible in magnitude and are statistically insignificant.

effect remained dominant across most of the distribution but reduced in magnitude, which is consistent with the analysis at the mean, suggesting that any remaining differences in wages were still primarily explained by characteristic difference in the profile of workers. The exceptions are at the very bottom and very top of the distribution where the structure effect is dominant. The sign of this effect is also now negative and its magnitude is larger than before, again consistent with the mean case. Additionally, the magnitude of the structure effect grows with wages, which indicates that higher-wage workers experienced larger changes in the returns to various individual-level characteristics compared to their lower-wage counterparts. Panel (c) shows that one additional year later, these dynamics with respect to the structure effect largely remained intact. As of 2022Q2, the structure effect was now dominant across most of the distribution, which is consist with the mean case, while the magnitude of most composition effect estimates were relatively small – reflecting the partial employment recovery thus far and hence more similar profile of workers compared to the pre-pandemic period. The opposite signs of the effects imply that the increase in wages due to any remaining differences in the profile of workers partially offset the larger decrease in wages brought on by changes in the associated returns to individual-level characteristics.

In Figure 20 we present the detailed decomposition of the full composition effects observed above across the wage distribution for each of the three periods of interest. As shown in panel (a), at the pandemic's onset, the five dominant covariates observed in the mean case above – trade union membership, main occupation, years of education, formal sector employment, and public sector employment – are evident across most of the wage distribution, however the magnitudes of their respective influences varies. Towards the bottom of the distribution, changes in the characteristics of workers with respect to education, formal sector employment, and unionisation primarily explain the composition effect here. Because the full structure effect is dominant at this part of the distribution as shown in the preceding figure, these characteristic changes only partially explain the increase in real wages at the bottom. Instead, changes in the returns to certain characteristics primarily do so, which is examined in more detail later. Towards the middle, education, unionisation (now to a greater extent relative to the bottom), formal sector employment (now to a lesser extent), occupation, and to a small extent public sector employment explain the composition effect. The full composition effect is dominant here, and hence changes in the characteristics of workers with respect to these characteristics primarily explain the increase in wages at this point of the distribution. Towards the top, education and occupation remain influential as well as public sector employment to a lesser extent, while unionisation reduces in importance. Importantly, the magnitude of the education coefficient is relatively constant and positive across the entire distribution, suggesting that changes in the education profile of the

employed population explained a similar absolute (but not relative[44]) amount of the increase in wages at the pandemic's onset regardless of the point of the wage distribution. Additionally, it is again notable that the composition effect is primarily explained not by demographics but instead by labour market characteristics, as in the mean case.

Figure 21: Recentered Influence Function detailed decomposition of composition effect across the wage distribution, by period



Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b;2021b; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Hourly wages adjusted for inflation and expressed in June 2022 Rands. Decomposition for categorical variables (industry, occupation, race, age, and province) based on "normalized" effects; that is, effects are expressed as deviation contrasts from the grand mean..

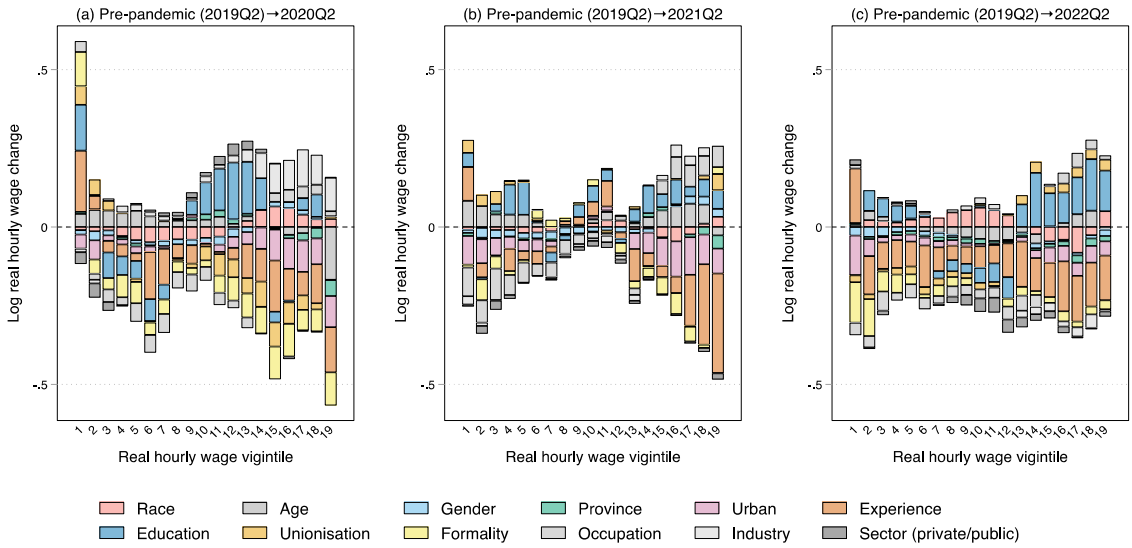
The estimates in panel (b) show that one year after the pandemic's onset, education remained an important contributor to the composition effect across the entire distribution, with the coefficient only marginally reducing in magnitude compared to the preceding period. This, in addition to union membership's continuing role, is consistent with the analysis at the mean. Unlike education however, union membership remains important in explaining the composition effect only towards the middle of the wage distribution and not at either the lower or upper tails, which is consistent with the preceding period. Towards the top, occupation increasingly explains the composition effect, as in the preceding period, but to a marginally-lower extent. This latter estimate is not uncovered in the mean analysis, which again highlights the advantage of the RIF approach. The magnitudes of the remaining covariates are all close to zero, reflecting the growing similarity of worker profiles relative to the pre-pandemic

[44] This is simply because the size of the composition effect varies across the wage distribution, and hence the relative share of the composition effect explained by a relatively constant education coefficient varies.

period as the labour market recovered. One additional year later, as shown in panel (c), education remained as an important and the dominant contributor to the composition effect across the entire distribution, which is consistent with the analysis at the mean, while occupation also remains important but again only towards the top. The magnitudes of the remaining covariates are negligible. This suggests that while characteristics of workers in 2022Q2 were more similar to those in the pre-pandemic period, some differences remained. Specifically, workers in the later period exhibited higher education levels and were more concentrated in higher-skilled occupations relative to those in the pre-pandemic period. However, because the full structure effect is dominant during this period, these differences in characteristics only partially offset the changes in the returns to various characteristics which drove real wages downwards at the end of the period.

In Figure 21 we present the detailed decomposition of the full structure effects observed above across the wage distribution for each of the three periods of interest. As in the mean case, the narrative pertaining to these detailed effects is less clear, primarily because the estimates for the majority of covariates are very small in magnitude and are statistically insignificantly different from zero. However, a few do stand out. As shown in panel (a), at the pandemic's onset, changes in the returns to education, experience, and formal sector employment placed upward pressure on wages at the bottom of the distribution – the only part of the distribution where the overall structure effect was dominant as shown above. The influence of these covariates are not evident when examining wage changes at the mean, which again highlights the existence of heterogeneities in the drivers of wage changes across the distribution. One year after the pandemic's onset, recall that the full, negative structure effects rise in importance from the middle to the top of the distribution but only partially offset the rise in wages driven by the dominant full composition effects. As shown in panel (b), from the middle to the top of the distribution changes in the associated returns to experience and residing in an urban area are largest in placing downward pressure on wages. This latter covariate is also evident in the mean case, but not the former. One additional year later, as shown in panel (c), the influence of the urban covariate observed for the period prior fell away while that of the experience covariate not only persisted spread to push downward pressure on wages across most of the distribution. Notably, this downward pressure from changes in the returns to experience at the top of the distribution was mostly offset by upward pressure induced by changes in the returns to education. Overall then, these estimates reveal a significant amount of heterogeneity in the drivers of the structure effect, both at a given part of the distribution, across the distribution, as well as over time as the pandemic progressed.

Figure 22: Recentered Influence Function detailed decomposition of structure effect across the wage distribution, by period



Author's own calculations. Source: QLFS 2019Q2, 2020Q2, 2021Q2, 2022Q2 (Statistics South Africa, 2019b; 2020b;2021b; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to those of working-age (15 to 64 years) employed. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Hourly wages adjusted for inflation and expressed in June 2022 Rands. Decomposition for categorical variables (industry, occupation, race, age, and province) based on "normalized" effects; that is, effects are expressed as deviation contrasts from the grand mean. Constant term omitted.

In summary, the decomposition analysis revealed a substantial degree of heterogeneity in the drivers of wage changes both at the mean and across the wage distribution over time. It is clear that a change in the characteristics of workers primarily explains the rise in wages both at the mean and across most of the distribution at the pandemic's onset, which itself is attributable to the notably regressive distribution of job loss observed prior. The specific characteristics which drive this change are trade union membership, main occupation, years of education, formal sector employment, and public sector employment, which are significant across the distribution but vary in influence. Importantly, changes in the returns to individual-level characteristics – specifically industry at the mean – played a more muted but non-negligible role in also driving wages upwards across most of the distribution at the pandemic's onset. The bottom of the distribution serves as the exception where changes in the returns to such characteristics, specifically education, experience, and formal sector employment, explain a greater share of the rise in wages at the pandemic's onset compared to the composition effect. As the pandemic progressed and employment partially recovered, the reductions in real wages across the distribution toward pre-pandemic levels were partially explained by the characteristics of workers more closely (but not completely) resembling those of the pre-pandemic period, but notably were primarily explained by persistent changes in the returns to various characteristics which vary across the wage distribution. This

latter funding is indicative of potential longer-term effects of the pandemic on the structure of the South African labour market.

6. Conclusion

In this paper, we conducted a micro-econometric analysis of the evolution of the level and nature of wage inequality and its drivers during the first two years of the COVID-19 pandemic in South Africa. To do so, we made use of nationally representative, individual-level, cross-sectional household survey data collected from 2019 to 2022, including raw, unimputed wage data provided by StatsSA not available in the public domain. A range of statistical measures and econometric techniques were employed to examine the quality of the data and answer three key research questions relating to the level and nature of South Africa's pre-pandemic wage distribution, changes to the wage distribution in response to the pandemic at its onset and over time, and the compositional and structural drivers of these changes.

First, we show that the missing wage data in the survey is both non-negligible in magnitude, with over a third of workers in the average wave do not report any wage information at all, and is non-randomly distributed, with non-response being highly inversely correlated with wages itself. These two characteristics justify an imputation procedure, however we provide evidence that the imputations in the public QLFS data are of very poor quality and that the use of either this data or the observed reported data alone results in an underestimation of wages across the entire distribution with estimates from the former exhibiting greater volatility over time, which has negative implications for any distributional analysis. We obtain reliable estimates of the wage distribution by adjusting the observed reported data for outliers using a parametric outlier detection model and missing data using an iterative, stochastic, and parametric imputation procedure which explicitly incorporates uncertainty inherent in the imputed values into the estimates, and thereafter conduct a battery of diagnostic tests to assess the quality of the imputations and sensitivity of the estimates.

Second, we find that wage inequality in the South African labour market was extremely high and stable in the year preceding the pandemic, regardless of measure. Using cross-sectional samples of the employed, at the pandemic's onset the distribution experienced a significant rightwards shift accompanied however by a very marginal change in its shape, indicating little to no change in wage inequality. As observed in other contexts, this rise in wages across the distribution was however mechanical and driven by a compositional change in the employed population induced by a regressive distribution of job loss. Workers at the bottom exhibiting job loss probabilities 3.5 times those of workers at the top, an outcome which we show may plausibly be explained by varying propensities to work in

'essential' jobs and work-from-home. Accounting for this selection using multiple methods, we calculate composition-controlled inequality indices which increase significantly at the pandemic's onset. Given that historical trends suggest inequality may have risen in the pandemic's absence, we estimate a counterfactual which suggests that approximately half of this rise – or up to 8 percent or 4.9 Gini points – is explained by the pandemic itself. Overall then, not accounting for such composition changes may lead to misinterpretations of wage inequality dynamics during this period. This rise in inequality was however transient, with wages quickly returning to their pre-pandemic levels in the period thereafter.

Third and finally, we show that the drivers of wage changes both at the mean and across the wage distribution were very heterogenous during a given period and over time. At the pandemic's onset, over 70 percent of the rise in the mean wage is explained by changes in the characteristics of workers, specifically with respect to five covariates: trade union membership, main occupation, years of education, formal sector employment, and public sector employment. This effect is dominant across most of the distribution and is consistent with prior findings, but the individual influences of the above covariates vary. Changes in the returns to characteristics – specifically industry at the mean – played a more muted but non-negligible role in driving wages across the distribution upwards. As the pandemic progressed and employment partially recovered, the reduction in real wages toward pre-pandemic levels was partially explained by a more (but not completely) similar profile of workers compared to the pre-pandemic period. However, persistent changes in the returns to various characteristics, which vary across the distribution, served as the dominant driver here, which is indicative of potential longer-term effects of the pandemic on the structure of the labour market.

Overall, it is clear that the COVID-19 pandemic and associated regulations had a significant effect on wages and wage inequality in the South African labour market. Although these effects largely appear to have been transient in nature, it is concerning that extreme levels of wage inequality persisted for at least two years following the pandemic's onset, despite partial but notable labour market recovery with respect to employment and working hours.

References

Abayomi, K., Gelman, A. and Levy, M., 2008. 'Diagnostics for multivariate imputations.' *Journal of the Royal Statistical Society*, Series C(57): 273–91.

Allison, P.D., 1978. 'Measures of Inequality.' *American Sociological Review*, 43(6): 865–80.

Atkinson, A.B., 1970. 'On the measurement of inequality.' *Journal of Economic Theory*, 2(3):244–63.

Atkinson, A.B. and Brandolini, A., 2010. 'On analyzing the world distribution of income.' *The World Bank Economic Review*, 24(1): 1-37.

Autor, D., Dube, A. and McGrew, A., 2023. 'The unexpected compression: Competition at work in the low wage labor market.' National Bureau of Economic Research Working Paper No. w31010. National Bureau of Economic Research.

Baker, M.G., 2020. 'Nonrelocatable Occupations at Increased Risk During Pandemics: United States, 2018.' *American Journal of Public Health,* 110: 1126-32.

Bhorat, H., Lilenstein, K., Oosthuizen, M. and Thornton, A., 2020. 'Wage polarization in a high-inequality emerging econoour: The case of South Africa.' WIDER Working Paper 2020/55. Helsinki: UNU-WIDER.

Bhorat, H., Lilenstein, A. and Stanwix, B., 2021. 'The Impact of the National Minimum Wage in South Africa: Early Quantitative Evidence.' Development Policy Research Unit Working Paper 202104. DPRU, University of Cape Town.

Bhorat, H., Stanwix, B. and Thornton, A., 2022. 'Changing dynamics in the South African labour market' In: The Oxford Handbook of the South African Econoour. Edited by: A. Oqubay, F Tregenna and we Valodia. Oxford University Press.

Blinder, A., 1973. 'Wage Discrimination: Reduced Form and Structural Estimates.' *Journal of Human Resources*, 8(4): 436–55.

Bollinger, C.R. and Hirsch, B.T., 2006. Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching. *Journal of Labor Economics*, 24(3): 483-519.

Cajner, T., Crane, L.D., Decker, R.A., Grigsby, J., Hamins-Puertolas, A., Hurst, E., Kurz, C. and Yildirmaz, A., 2020. 'The US labor market during the beginning of the pandemic recession.' National Bureau of Economic Research Working Paper No. w27159. National Bureau of Economic Research.

Casale, D. and Shepherd, D., 2021. 'The gendered effects of the COVID-19 crisis and ongoing lockdown in South Africa: Evidence from NIDS-CRAM Waves 1- 5.' National Income Dynamics Study – Coronavirus Rapid Mobile Survey Wave 5 Policy Paper 3. Available here: https://cramsurvey.org/wp-content/uploads/2021/07/3.-Casale-D.- -Shepherd-D.-2021-The-gendered-effects-of-the-Covid-19-crisis-and-ongoing-lockdown-in-South-Africa-Evidence-from-NIDS-CRAM-Waves-1-–-5..pdf.

Cornia, G.A., 2014. 'Inequality trends and their determinants: Latin America over the period 1990-2010' In: Falling inequality in Latin America. Edited by G.A. Cornia. Oxford University Press.

Cowell, F.A., 2011. Measuring inequality. Oxford University Press: New York.

Cribb, J., Waters, T., Wernham, T. and Xu, X., 2021. 'The labour market during the pandemic.' Living standards, poverty and inequality in the UK. United Kingdom: The Institute for Fiscal Studies.

Díaz Pabón, F.A., Leibbrandt, M., Ranchhod, V. and Savage, M., 2021. 'Piketty comes to South Africa.' *The British Journal of Sociology*, 72(1): 106-24.

Dingel, J.we. and Neiman, B., 2020. 'How many jobs can be done at home?.' *Journal of Public Economics*, 189: 104235.

Finn, A., Leibbrandt, M. and Ranchhod, V., 2016. 'Patterns of persistence: Intergenerational mobility and education in South Africa.' Cape Town: SALDRU, University of Cape Town. SALDRU Working Paper Number 175/ NIDS Discussion Paper 2016/2.

Finn, A. and Leibbrandt, M., 2018. 'The evolution and determination of earnings inequality in Post-Apartheid South Africa.' WIDER Working Paper 2018/83. Helsinki: UNU-WIDER.

Firpo, S., Fortin, N. and Lemieux, T., 2009. 'Unconditional Quantile Regressions.' *Econometrica*, 77(3): 953–73.

Firpo, S., Fortin, N. and Lemieux, T., 2018. 'Decomposing Wage Distributions Using Recentered Influence Function Regressions.' *Econometrics*, 6(28): 1-40.

Fortin, N., Lemieux, T. and Firpo, S., 2011. 'Decomposition Methods in Economics.' In O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Volume 4A. Amsterdam: Elsevier.

Furceri, D., Loungani, P., Ostry, J.D. and Pizzuto, P., 2022. 'Will COVID-19 have long-lasting effects on inequality? Evidence from past pandemics.' *The Journal of Economic Inequality*, 20: 811–39.

Hill, R. and Köhler, T., 2021. 'Mind the gap: Analysing the effects of South Africa's national lockdown on gender wage inequality.' National Income Dynamics Study – Coronavirus Rapid Mobile Survey Wave 2 Policy Paper 7. Available here: https://cramsurvey.org/wp-content/uploads/2020/09/7.-Hill-R.-_-Köhler-T.-2020-Mind-the-gap-Analysing-the-effects-of-South-Africa's-national-lockdown-on-gender-wage-inequality.pdf.

Kerr, A., 2021. 'Measuring earnings inequality in South Africa using household survey and administrative tax microdata'. WIDER Working Paper No. 2021/82. Helsinki: United Nations World Institute for Development Economic Research (UNU-WIDER).

Kerr, A., 2022. 'The PALMS and the ESES projects. Earnings in the Quarterly Labour Force Survey (QLFS)'. Unpublished presentation. 15 years of the Data Quality Programme. DataFirst: University of Cape Town.

Kerr, A. and Thornton, A., 2020. 'Essential workers, working from home and job loss vulnerability in South Africa.' DataFirst Technical Paper No. 41. DataFirst, University of Cape Town.

Kerr, A. and Wittenberg, M., 2019a. 'Earnings and employment microdata in South Africa'. WIDER Working Paper No. 2019/47. Helsinki: United Nations World Institute for Development Economic Research (UNU-WIDER).

Kerr, A. and Wittenberg, M., 2019b. 'A Guide to version 3.3 of the Post-Apartheid Labour Market Series'. Available at: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/434/download/8243.

Kerr, A. and Wittenberg, M., 2021. 'Union wage premia and wage inequality in South Africa.' *Economic Modelling*, 97: 255–71.

Khanyile, S. and Kerr, A. 2022. 'The Impact of the Quarterly Labour Force Survey (QLFS) Earnings Imputations'. Unpublished presentation. University of Cape Town School of Economics Departmental Seminar. University of Cape Town.

Köhler, T., 2023. 'What we know about COVID-19 and the South African labour market.' Research on Socio-Economic Policy (RESEP) Research Note. University of Stellenbosch.

Köhler, T., Bhorat, H., Hill, R. and Stanwix, B. 2023a. 'Lockdown stringency and employment formality: evidence from the COVID-19 pandemic in South Africa.' *Journal for Labour Market Research,* 57(3): 1-28.

Köhler, T., Bhorat, H., and Hill, R. 2023b. 'Wage Subsidies and Job Retention in a Developing Country: Evidence from South Africa.' Development Policy Research Unit Working Paper 202302. DPRU, University of Cape Town.

Leibbrandt, M., Finn, A. and Woolard, we., 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., 2021. 'South Africa: The Top End, Labour Markets, Fiscal Redistribution, and the Persistence of Very High Inequality' In: Inequality in the Developing World. Edited by: C. Gradín, M. Leibbrandt, and F. Tarp. Oxford University Press, United Nations University World Institute for Development Economics Research (UNU-WIDER).

Leibbrandt, M. and Díaz Pabón, F.A., 2022. 'Inequality in South Africa' In: The Oxford Handbook of the South African Econoour. Edited by: A. Oqubay, F Tregenna and we Valodia. Oxford University Press.

Lemieux, T., 2006. The "Mincer Equation" Thirty Years After *Schooling, Experience, and Earnings*. In: Grossbard, S. (eds) Jacob Mincer A Pioneer of Modern Labor Economics. Springer, Boston.

Martin, M.A., Lennon, R.P., Smith, R.A., Myrick, J.G., Small, M.L., Van Scoy, L.J., 2022. 'Essential and non-essential US workers' health behaviors during the COVID-19 pandemic.' *Preventive Medicine Reports*, 29: 101889.

McGregor, T., Smith, B. and Willis, S., 2019. 'Measuring inequality.' *Oxford Review of Economic Policy*, 35(3): 368–95.

Mincer, J., 1974. Schooling, Experience and Earnings. Columbia University Press: New York.

Mongey, S. and Weinberg, A., 2020. 'Characteristics of workers in low work-from-home and high personal-proximity occupations.' White paper. Chicago, IL: University of Chicago, Becker Friedman Institute for Economics.

Montenovo, L., Jiang, X., Lozano-Rojas, F., Schmutte, we., Simon, K., Weinberg, B.A. and Wing, C., 2022. 'Determinants of Disparities in Early COVID-19 Job Losses.' *Demography,* 59(3): 827–55.

Oaxaca, R., 1973. 'Male–Female Wage Differentials in Urban Labor Markets.' *International Economic Review*, 14(3): 693–709.

Oaxaca, R. and Sierminska, E., 2023. 'Oaxaca-Blinder Meets Kitagawa: What is the Link?.' IZA Discussion Paper No. 16188. Institute of Labor Economics.

Patrinos, H.A., 2016. 'Estimating the return to schooling using the Mincer equation.' *IZA World of Labor*, 278.

Piraino, P., 2015. 'Intergenerational earnings mobility and equality of opportunity in South Africa.' *World Development*, 67: 396-405.

Ranchhod, V. and Daniels, R., 2021. 'Labour Market Dynamics in South Africa at the Onset of the COVID-19 Pandemic.' *South African Journal of Economics*, 89(1): 44-62.

Rubin, D.B., 1987. *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Sen, A., 1997. *On economic inequality*. Oxford University Press.

Shifa, M. and Ranchhod, V., 2019. Handbook on Inequality Measurement for Country Studies. Cape Town: African Centre of Excellence for Inequality Research, University of Cape Town.
Shifa, M., Mabhena, R., Ranchhod, V. and Leibbrandt, M., 2023. 'An assessment of inequality estimates for the case of South Africa'. WIDER Working Paper No. 2023-90. Helsinki: United Nations World Institute for Development Economic Research (UNU-WIDER).

Shorrocks, A.F., 1984. 'Inequality decomposition by population subgroups.' *Econometrica: Journal of the Econometric Society*: 1369-1385.

Statistics South Africa, 2019a, 'Quarterly Labour Force Survey (2019Q1).' Dataset. Pretoria: Statistics South Africa.

———, 2019b, 'Quarterly Labour Force Survey (2019Q2).' Dataset. Pretoria: Statistics South Africa.

———, 2019c, 'Quarterly Labour Force Survey (2019Q3).' Dataset. Pretoria: Statistics South Africa.

———, 2019d, 'Quarterly Labour Force Survey (2019Q4).' Dataset. Pretoria: Statistics South Africa.

———, 2020a, 'Quarterly Labour Force Survey (2020Q1).' Dataset. Pretoria: Statistics South Africa.

———, 2020b, 'Quarterly Labour Force Survey (2020Q2).' Dataset. Pretoria: Statistics South Africa.

———, 2020c, 'Quarterly Labour Force Survey (2020Q3).' Dataset. Pretoria: Statistics South Africa.

———, 2020d, 'Quarterly Labour Force Survey (2020Q4).' Dataset. Pretoria: Statistics South Africa.

———, 2021a, 'Quarterly Labour Force Survey (2021Q1).' Dataset. Pretoria: Statistics South Africa.

———, 2021b, 'Quarterly Labour Force Survey (2021Q2).' Dataset. Pretoria: Statistics South Africa.

———, 2021c, 'Quarterly Labour Force Survey (2021Q3).' Dataset. Pretoria: Statistics South Africa.

———, 2021d, 'Quarterly Labour Force Survey (2021Q4).' Dataset. Pretoria: Statistics South Africa.

———, 2022a, 'Quarterly Labour Force Survey (2022Q1).' Dataset. Pretoria: Statistics South Africa.

———, 2022b, 'Quarterly Labour Force Survey (2022Q2).' Dataset. Pretoria: Statistics South Africa.

———, 2022c, 'Consumer Price Index: June 2022'. Statistical release P0141. Pretoria: Statistics South Africa. Available here: https://www.statssa.gov.za/publications/P0141/P0141June2022.pdf.

Stevens, J.P., 1984. 'Outliers and influential data points in regression analysis.' *Psychological bulletin*, 95(2): 334.

Wittenberg, M., 2017. 'Wages and wage inequality in South Africa 1994-2011: Part 2 – Inequality measurement and trends.' *South African Journal of Economics*, 85(2): 298-318.

Wittenberg, M., 2018. 'The Top Tail of South Africa's Earnings Distribution 1993–2014: Evidence from the Pareto Distribution.' SALDRU Working Paper 224. Cape Town: SALDRU, University of Cape Town.

Appendix

Table A1: Balance table of observable covariates by wage reporting status, 2020Q1

	(1) Reported exact wage	(2) Reported bracket only	(3) Reported neither	(4) Difference (1)- (2)	(5) Difference (1)- (3)
Age (years)	38.89	40.14	39.73	-1.24***	-0.84***
	(0.13)	(0.21)	(0.19)	(0.25)	(0.24)
Female	0.46	0.44	0.42	0.02**	0.04***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Years of education	10.24	11.92	11.90	-1.69***	-1.66***
	(0.05)	(0.07)	(0.05)	(0.08)	(0.07)
African/Black	0.85	0.72	0.63	0.12***	0.22***
	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)
Urban	0.69	0.76	0.88	-0.08***	-0.19***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Informal sector	0.22	0.15	0.15	0.07***	0.08***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Public sector	0.15	0.24	0.17	-0.09***	-0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Union member	0.24	0.40	0.31	-0.16***	-0.07***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Author's own calculations. Source: QLFS 2020Q1 (Statistics South Africa 2020a).
*Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. All estimates are weighted using the sample weights. Wave-specific mean for trade union membership is assigned to individuals with missing trade union data. Standard errors are adjusted for the complex survey design and are presented in parentheses. * p<0.10; ** p<0.050; *** p<0.001.*

Table A2: Linear probability model estimates of the correlates of having missing wage data

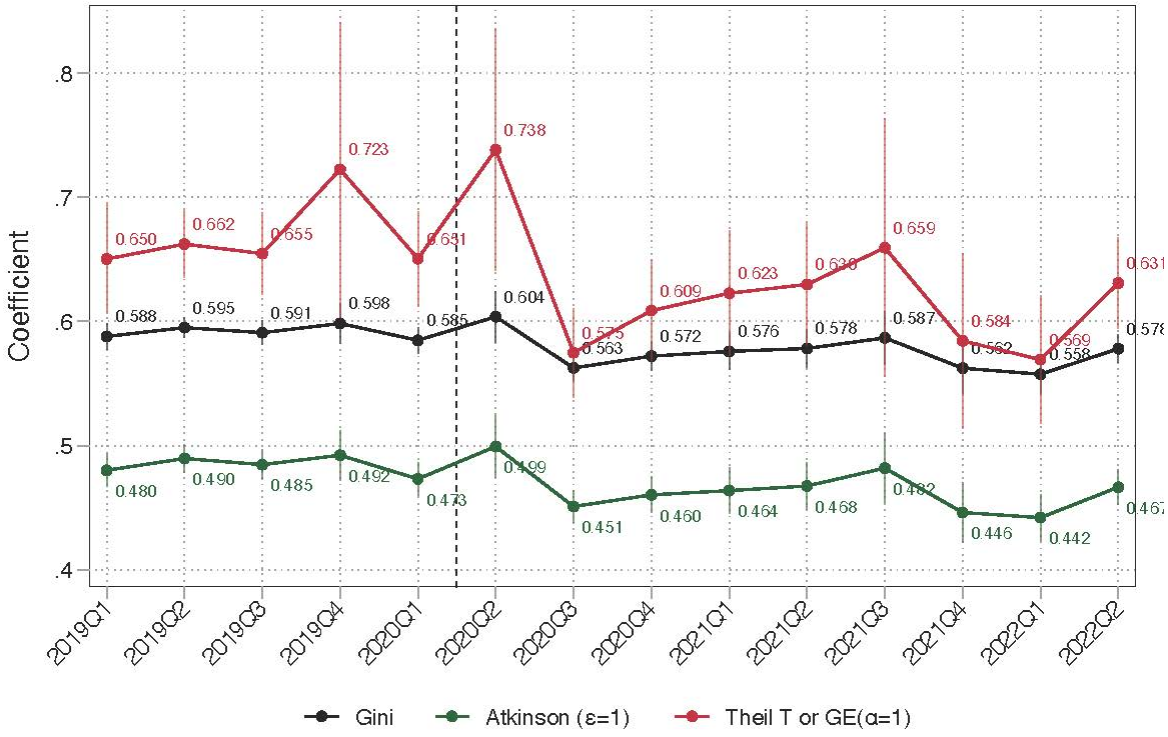
Outcome variable:	(1) =1 if missing exact value and bracket =0 if reported exact value or bracket	(2) =1 if missing exact value and bracket =0 if reported exact value	(3) =1 if missing exact value only =0 if reported exact value
Wage interval (base = monthly)			
Weekly	-0.100***	-0.110***	-0.135***
	(0.004)	(0.004)	(0.004)
Fortnightly	-0.092***	-0.094***	-0.125***
	(0.006)	(0.005)	(0.006)
Daily	-0.088***	-0.110***	-0.174***
	(0.005)	(0.004)	(0.004)
Hourly	-0.150***	-0.193***	-0.239***
	(0.011)	(0.010)	(0.011)
Annually	-0.163***	-0.198***	-0.124***
	(0.021)	(0.021)	(0.020)
Refusal/DK	0.378***	0.670***	0.592***
	(0.003)	(0.003)	(0.002)
Age (years)	-0.002***	-0.002**	0.000
	(0.001)	(0.001)	(0.001)
Age squared	0.000**	0.000*	0.000
	(0.000	(0.000	(0.000
Female	-0.017***	-0.027***	-0.030***
	(0.002)	(0.002)	(0.002)
Years of schooling	0.005***	0.010***	0.013***
	(0.001)	(0.001)	(0.001)

Race (base = African/Black)			
Coloured	0.044***	0.080***	0.085***
	(0.004)	(0.004)	(0.004)
Indian/Asian	0.126***	0.098***	0.069***
	(0.006)	(0.006)	(0.006)
White	0.097***	0.112***	0.110***
	(0.004)	(0.004)	(0.004)
Province (base = WC)			
EC	0.000	0.092***	0.115***
	(0.004)	(0.004)	(0.004)
NC	0.019***	0.083***	0.099***
	(0.006)	(0.006)	(0.006)
FS	-0.064***	0.026***	0.087***
	(0.005)	(0.005)	(0.005)
KZN	-0.018***	0.118***	0.174***
	(0.004)	(0.004)	(0.004)
NW	-0.143***	-0.022***	0.027***
	(0.005)	(0.006)	(0.005)
GP	0.122***	0.246***	0.278***
	(0.004)	(0.004)	(0.004)
MP	0.087***	0.219***	0.261***
	(0.005)	(0.005)	(0.005)
LP	-0.205***	-0.006	0.047***
	(0.005)	(0.005)	(0.005)
One-digit industry (base = agriculture)			
Mining	0.074***	0.103***	0.132***
	(0.008)	(0.008)	(0.008)
Manufacturing	0.083***	0.090***	0.094***
	(0.006)	(0.006)	(0.005)
Utilities	0.110***	0.103***	0.079***
	(0.013)	(0.013)	(0.012)
Construction	0.064***	0.067***	0.073***
	(0.006)	(0.006)	(0.006)
Trade	0.071***	0.070***	0.078***
	(0.005)	(0.005)	(0.005)
Transport	0.101***	0.093***	0.094***
	(0.006)	(0.006)	(0.006)
Finance	0.063***	0.054***	0.057***
	(0.005)	(0.005)	(0.005)
Community and social services	0.042***	0.042***	0.052***
	(0.006)	(0.006)	(0.005)
Private households	0.015*	0.002	0.000
	(0.009)	(0.008)	(0.008)
One-digit occupation (base = manager)			
Professional	-0.032***	-0.016***	0.000
	(0.006)	(0.006)	(0.005)
Technician	-0.001	0.004	0.007
	(0.005)	(0.005)	(0.005)
Clerk	0.003	0.005	0.001
	(0.005)	(0.005)	(0.005)
Sales and services	-0.056***	-0.077***	-0.076***
	(0.005)	(0.005)	(0.004)
Skilled agriculture	-0.052***	-0.073***	-0.090***
	(0.015)	(0.015)	(0.015)
Craft	-0.031***	-0.048***	-0.047***
	(0.005)	(0.005)	(0.005)
Plant and machine operator	-0.023***	-0.042***	-0.046***
	(0.005)	(0.006)	(0.005)
Elementary	-0.064***	-0.090***	-0.100***
	(0.005)	(0.005)	(0.004)
Domestic worker	-0.066***	-0.091***	-0.113***

	(0.010)	(0.009)	(0.009)
Urban	0.065***	0.073***	0.050***
	(0.003)	(0.003)	(0.003)
Public sector	-0.022***	-0.007*	-0.006*
	(0.004)	(0.004)	(0.003)
Informal sector	-0.004	-0.006*	-0.016***
	(0.003)	(0.003)	(0.003)
Trade union membership	-0.001	0.049***	0.081***
	(0.003)	(0.003)	(0.003)
Constant	0.132***	0.010	0.030
	(0.022)	(0.022)	(0.021)
Observations	177 183	141 083	177 183
R^2	0.250	0.461	0.406

Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
*Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed. Unweighted estimates presented. All models control for wave fixed effects. Binary indicator for missing trade union data included as a covariate, while the wave-specific mean for trade union membership is assigned to individuals with missing trade union data. Standard errors presented in parentheses. * p<0.10; ** p<0.050; *** p<0.001.*

Figure A1: Relative wage inequality estimates by measure excluding furloughed workers, 2019Q1 – 2022Q2



Author's own calculations. Source: QLFS 2019Q1 - 2022Q2 (Statistics South Africa, 2019a; 2019b; 2019c; 2019d; 2020a; 2020b; 2020c; 2020d; 2021a; 2021b; 2021c; 2021d; 2022a; 2022b).
Notes: Unimputed wage data provided by StatsSA. Sample restricted to the working-age (15 to 64 years) employed who reported working non-zero hours. Estimates are weighted using sampling weights. Standard errors are adjusted for the complex survey design. Spikes represent 95 percent confidence intervals.

Development Policy Research Unit
University of Cape Town
Private Bag, Rondebosch 7701
Cape Town, South Africa
Tel: +27 21 650 5701
www.dpru.uct.ac.za