SA-TIED

Southern Africa – Towards Inclusive Economic Development

WORKING PAPER 219

The gender wage gap in South Africa

Insights from administrative tax data

Michelle Pleace, Matthew Clance, and Nicky Nicholls *

October 2023

About the project

Southern Africa – Towards Inclusive Economic Development (SA-TIED)

SA-TIED is a unique collaboration between local and international research institutes and the government of South Africa. Its primary goal is to improve the interface between research and policy by producing cutting-edge research for inclusive growth and economic transformation in the southern African region. It is hoped that the SA-TIED programme will lead to greater institutional and individual capacities, improve database management and data analysis, and provide research outputs that assist in the formulation of evidence-based economic policy.

The collaboration is between the United Nations University World Institute for Development Economics Research (UNU-WIDER), the National Treasury of South Africa, the South African Revenue Services, and other universities and institutes. It is funded by the National Treasury of South Africa, the Delegation of the European Union to South Africa and UNU-WIDER through the Institute's contributions from Finland, Sweden, and the United Kingdom to its research programme.

* Department of Economics, University of Pretoria, Pretoria, South Africa, michelle.pleace@up.ac.za

The gender wage gap in South Africa

Insights from administrative tax data

Michelle Pleace, Matthew Clance, and Nicky Nicholls*

October 2023

Abstract: The attainment of financial independence by women holds significant importance for women's empowerment and has implications for the increasing prevalence of female-headed households in the country. Previous estimates of the gender wage gap in South Africa have relied on self-reported surveys, which may not accurately represent the income distribution due to underrepresentation of the upper tail and potential misreporting. To address these limitations, we use administrative tax data to estimate the income differential by gender within the formal economy from 2008 to 2021. Our findings reveal a widening gender wage gap over time. In 2008, women earned approximately 89 cents for every ZAR1 earned by men, whereas by 2021, this ratio decreased to 78 cents. Notably, the highest degree of inequality by gender is observed in the lower tail of the income distribution. However, when examining the high-skill sector exclusively, the income differential is largest at the 90th percentile of the conditional income distribution. Furthermore, within this high-skill industry, women are more likely to occupy lower-paid positions. Our findings suggest that efforts focusing on low-income categories and improving the representation of women in senior management might reduce the income disparity by gender.

Key words: gender wage gap, formal economy, South Africa, tax data, income distribution, gender

JEL classification: J16, J30

Acknowledgements: The authors report there are no competing interests to declare.

* Department of Economics, University of Pretoria, Pretoria, South Africa, michelle.pleace@up.ac.za

1 Introduction

Female autonomy is necessary to overcome entrenched social norms that restrict female labour to domestic tasks (Engels 1972). Financial independence enhances the agency of women (Anderson and Eswaran 2009) and is correlated with better child survival rates (Eswaran 2002). Reducing the gender wage gap is particularly important for female-headed households, which account for 42.1% of homes in South Africa (StatsSA 2022). As long as gender-based income disparities persist, individuals living in female-headed households are more likely to experience poverty compared to those in male-headed households (Mosomi 2019), influencing the welfare of the next generation. Reducing the gender wage differential enhances female welfare and empowerment as well as the welfare of the female-headed household (Danquah et al. 2021). We estimate the gender wage differential in the formal economy.

Previous studies have estimated the gender wage gap in South Africa using self-reported surveys. However, self-reported measures can be flawed in accurately capturing income (Pedace 2000; Bargain et al. 2018), and these measures may not represent the upper tail of the income distribution (Fortin et al. 2017). By utilizing administrative tax data, we include all employees working in the formal economy to estimate the gender wage gap. Developed countries have calculated the wage differential by gender using administrative data (Bayard et al. 2004; Hospido and Moral-Benito 2016; Fortin et al. 2017). A search retrieves only one case of administrative data being used to estimate the gender pay gap in the developing country context—that being Uruguay (Burdín et al. 2022).

Descriptive statistics show that the proportion of females in formal employment in South Africa has consistently increased since 2008. Figure 1 shows that an employment gap still remains; it decreased by 5.96 percentage points as of 2021. However, it is concerning that the income disparity at the median level has widened over the same period. In 2008, women earned 89 cents for every ZAR1 of men's income, but by 2021, this figure dropped to 78 cents. This is due in part to the faster growth of male median income relative to female median income. Additionally, we note a decline in the proportion of women in the top-income-earning category within the high-skill sector from 2015 to 2021.

Using linear regression analysis, we note that males earn significantly more income than females at the mean of the income distribution. Further, we find that the 'poorest 10%' income category is the most unequal by gender. Of note, the administrative data do not capture informal activities, meaning that we cannot accurately capture the gender pay gap at the lowest end of the income distribution. When limiting the sample to the high-skill sector, the gender wage differential is smaller in the high-skill sector than in the other industries. Also, with the use of quantile regression analysis, we note that the gender wage differential is greatest in the 'richest 10%' income category, possibly due to the low number of women in management positions (StatsSA 2022).

Section 2 provides a review of the related literature, followed by Section 3 which outlines the methodology. In Section 4, descriptive results are presented, while Section 5 explains the empirical strategy. Section 6 presents the regression results, and Section 7 concludes.

2 Related literature

2.1 Women's financial freedom in developing countries

Zawaira et al. (2022) find that female labour force participation is lower in patriarchal societies than in matriarchal societies. Further, the authors find that gender attitudes can reduce female labour force participation, as the traditional gender roles of males being the income earners and females being caregivers persist across generations. In South Africa, Casale and Posel (2002) note that more women entered the formal and informal labour market during 1995–99 as women received less income support from men. This shows the dual responsibilities of women consisting of caregiving and earning income to support their families. Post-1999, the female labour force participation in South Africa continued to increase (ILO 2022), alluding to the importance of females earning their own income and being more financially independent.

Hanmer and Klugman (2016) find that in developing countries, four of five women experience at least one constraint to their freedom. Constraints include encountering intimate partner violence, restricted movement, lack of control over household resources, no control over contraception, and child marriage. The authors note that financial freedom plays an integral part in improving the agency of women.

Reducing the wage differential by gender is a significant step in empowering women to be financially independent, which is in turn related to improved well-being measures such as child nutrition and improved family health care (Donald et al. 2020). Therefore, it is important to determine the progress that women have made in South Africa's formal economy.

2.2 The income differential by gender in South Africa

The gender wage gap has been estimated using survey data (e.g., Ntuli 2014; Bhorat and Goga 2012; Mosomi 2019). A significant wage differential is found using the Labour Force Survey (LFS), October Household Survey (OHS), and Post-Apartheid Labour Market Series (PALMS). The surveys obtain income information by asking individuals whether they earn weekly/monthly/annual income between different bounds; therefore, a specific income amount is not disclosed in the surveys.

Investigating the Black South African population specifically, Ntuli (2014) conducts a longitudinal study (1995–2004), finding that the gender wage gap is the largest in the lowest-earning-income groups. This result is confirmed when including all race categories (Bhorat and Goga 2013) and over a longer period (1993–2019) (Mosomi 2019). Ntuli (2014) finds that the gender wage differential decreased from 1995 to 2004 in the 10th and 25th percentile groups using the LFS and OHS, which might be due to the implementation of the national minimum wage policy (Mosomi 2019). While the gender wage gap has improved at the lower tail, the overall trend of the median income gap remained unchanged from 1993 to 2019 (Mosomi 2019).

Other research has split the wage differential into explained and unexplained components. Grün (2004) implements the Oaxaca-Blinder (OB) decomposition to estimate the gender wage differential and finds that gender discrimination plays a role in the income gap.[1]

2.3 Explanations for the wage differential by gender

A strand of literature investigates the relationship between the income differential by gender and gender segregation within sectors. Roberts and Schöer (2021) find that females earn more in male-dominated industries relative to female peers in female-dominated occupations.[2] Landman and O'Clery (2020) attribute this to increasing female wages in order to retain female workers in male-dominated sectors so that these sectors comply with the South African Employment Equity Act.

---

[1] This decomposition estimates the difference in the mean wages by gender to the difference in mean values of the independent variables, which can be observable (education, age, occupation title) and unobservable (discrimination) variables.

[2] Roberts and Schöer (2021: 2) define a male-dominated industry as an industry that employs 66% males. This definition is against the usual industry dominance definition used in the literature, as the authors state, 'In the crossover literature, this cut-off for male-dominated ranges from 70 to 75% for the sectors that business-owners work in'.

Existing research has offered some reasons for why women earn less than men. Travis (2015) argues that the gap is due to the type of occupation that females enter and that females do not negotiate their salary to a level comparable with their male peers.

Fransen et al. (2012) claim that women do not have the required education or work experience to earn the same wages as men. In Europe, Kunze (2017) finds that the wage differential remains unchanged even as women increase their human capital to match a man's education level. Blau and Kahn (2017) analyse the trends of the wage differential and find that, after 2010, investment in human capital (education and experience) explains a small portion of the wage gap relative to the difference in occupations.

Occupational segregation contributes to the gender pay gap. Gradín (2021) finds that although women have access to higher-paying positions due to improved educational attainment, men continue to dominate managerial positions. Employers may prefer hiring males over equally performing females and might therefore be willing to pay an income premium for male workers rather than hiring a female (Bhorat and Goga 2012). Moss-Racusin et al. (2012) showed that hiring committees have a preference for male candidates even when an identical CV is used for female candidates. Further, there is a preference for male applicants to be appointed instead of female applicants until females are described with traditionally male adjectives such as assertive and competent (Rice and Barth 2016).

Lastly, in Africa, obstacles such as gender norms, patriarchal societies, and lack of opportunities can hinder females from obtaining tertiary education in male-dominated fields (Mukhwana et al. 2020).

2.4 Administrative data to observe a wage differential

Administrative data are beneficial over self-reported survey data. These data are nationally representative of the formal economy where survey data can undersample top income earners (Fortin et al. 2017), limiting the conclusions drawn using survey data (Bayard et al. 2004). Individuals in the low tail of the income distribution often over-report their income in surveys, while those in the upper tail of the distribution tend to self-report lower wages than their actual income (Pedace 2000).

Using administrative tax data, Fortin et al. (2017) explore the gender wage differential in the top 10% income distribution across Canada, Sweden, and the United Kingdom. The authors found that the gender wage differential improved between 1983 and 2015 for most women except those in the richest 0.1% of the income distribution. In France, Bargain et al. (2018) uses administrative tax records to identify the wage differential between public and private sector workers. The gender wage differential is highest at the top end of the income distribution in the private sector.[3]

Burdín et al. (2022) finds that inequality reduces significantly in Uruguay between 2009 and 2016 using self-reported income from household survey data. However, when using actual income from administrative data to analyse inequality over the same time period, inequality in Uruguay increases. This highlights the misreporting issue of self-report data when considering income.

In this paper, we use administrative tax data to obtain a recent estimate of the gender wage differential in the formal economy over time.

[3] The authors have access to age, gender, and type of occupation. They argue that fixed effects can account for other characteristics of the worker that are broadly time invariant, such as education levels, which are not present in administrative data. Therefore, we also use individual fixed effects to account for these differences in the population.

3 Data source and descriptive statistics

3.1 Data

Our study uses the South African Revenue Service administrative tax data for 2008–21 (National Treasury and UNU-WIDER 2022). The data set is constructed from anonymized individual tax records relating to income.

Labour income tax certificates

The IRP5 and IT3(a) are two tax certificates related to labour income that companies can issue to their employees. Companies must register as pay-as-you-earn entities unless none of their employees earn enough income to be liable for income tax. Only registered companies issue tax certificates (IRP5 or IT3(a)) to their employees. Individuals earning labour income may be subject to three types of tax: i) Pay As You Earn (PAYE) income tax, with a minimum tax threshold of ZAR83,100 in 2021. This means that individuals earning less than ZAR6,925 per month will not be required to pay income tax. ii) the Skills Development Levy (SDL) is a levy aimed at promoting learning and development of the workforce. Individuals contribute to this levy if they work 24 hours or more in a month. iii) The Unemployment Insurance Fund (UIF) provides short-term relief to individuals who become unemployed. Individuals contribute to this fund if they work 24 hours or more in a month.

Employers issue an IT3(a) certificate on behalf of workers who earn more than ZAR2,000 per month but are not liable for tax (including income tax, UIF, and SDL), likely because the employee works less than 24 hours per month.

IRP5 certificates are issued when workers earn more than ZAR2,000 per month but less than the minimum tax threshold and work more than 24 hours per month. In this case, individuals need to contribute to UIF and SDL. IRP5 certificates are also issued when individuals earn more than the tax threshold. In such cases, individuals contribute to income tax, UIF, and SDL.

Tax certificates list all income source codes that an employee earned during the tax year, as well as the number of periods that the individual worked for a registered firm in the formal economy. Each registered employer files a tax record for each employee who receives compensation. If a person works for multiple firms in one tax year, they will receive multiple tax certificates. As the firm issues these tax certificates to their workers, the data consist of a matched panel between firms and employees. This data set is known as the 'IRP5/IT3(a) data set.'

In summary, we have labour income data that are disclosed on an anonymized tax certificate for employees earning more than ZAR2,000 per month from registered entities in the formal economy. These data do not account for any informal activities, and personal characteristics of an individual (except gender, age, and the province that the individual lives in) are not disclosed on a tax certificate, such as the education of an individual.

Personal income tax certificate

A personal income tax certificate is known as the Income Tax Return (ITR12), where individuals disclose all income received from each income source code and report any deductions claimed for a specific tax year. Since individuals submit only one ITR12 tax certificate per year, income by source codes is not disaggregated by different employers. This data set is referred to as the 'ITR12 data set'.

An individual is not required to submit an ITR12 tax certificate unless three conditions are met: i) The individual earned labour income from a single employer, with a gross income of less than ZAR500,001 for the 2021 tax year, and did not have any additional sources of income during this period. ii) The individual did not receive any allowances from their employer, such as car or travel allowances. iii) The individual did not claim any tax-related deductions or rebates.

It is possible for individuals to have an IRP5/IT3(a) tax certificate and be exempt from submitting an ITR12 tax certificate due to the above exclusion criteria. Similarly, it is possible that individuals submit an ITR12 tax certificate, but they do not have an IRP5/IT3(a) certificate. For example, entrepreneurs who operate their own businesses do not submit an IRP5 certificate since they are not employees of the business. Instead, they submit an ITR12 tax certificate and disclose all income obtained from their business.

In summary, we can identify aggregate labour income earned by an individual in a tax year from a personal tax certificate. These data do not include any personal characteristics of the individual other than gender, age, and the province where the individual lives.

3.2 Data creation

To observe the gender wage gap in the formal economy, we consider labour income only. Kerr (2020) created a variable within the IRP5/IT3(a) data set to identify all labour income earned by individuals. The labour income variable includes salaries/wages, commission, allowances,[4] medical contributions, bursaries, and scholarships.[5] The labour income variable aligns well with the ‘compensation to employees’ estimate calculated in the National Income and Product Accounts by the South African Reserve Bank, illustrating that this variable is a good estimate for national income (Kerr 2020). Therefore, we use the labour income variable derived by Kerr (2020) as our income estimate. This labour income is deflated by the consumer price index to create a real income variable used in our analysis.[6] The IRP5/IT3(a) and ITR12 data sets have a South African rand amount that is associated with an income source code. Therefore, we are able to isolate labour income source codes as identified by Kerr (2020) to obtain the total labour income earned from different source codes by an individual in a year.

In the IRP5/IT3(a) data set, individuals can work for multiple firms where they would have multiple IRP5 certificates. As we do not include firm-level analysis, we sum all labour income earned by an individual for each year to obtain a total labour income estimate. In our data, an individual appears once a year with a corresponding total labour income.

The ITR12 data set captures personal wealth; therefore, individuals submit one ITR12 form annually. The ITR12 data disclose the income earned from all income source codes for each individual. We add all income from source codes identified as labour income by Kerr (2020). In other words, we have the total labour income disclosed on the ITR12 tax certificate for every individual.

We merge the IRP5/IT3(a) and ITR12 data sets together using an individual’s anonymized identification number. We keep individuals that have labour income disclosed from the IRP5/IT3(a) tax certificates and individuals that have labour income disclosed on the ITR12 tax certificate if this individual does not have an IRP5/IT3(a) tax certificate. Therefore, in our final data set, we have labour-income earners from

[4] This includes allowances for travel, uniforms, tools, computer, phones, meals, and accommodation.

[5] The FeesMustFall movement started in 2015, a student-led protest to increase government funding to public universities. We investigated whether there was an increase in the number of scholarships awarded to students, which would increase the value of the labour income variable estimated by Kerr (2020). In the Appendix, Table C5 shows the number of scholarships awarded every year. We do not see a sharp increase in the number of scholarships made available to students directly after 2015 from the FeesMustFall movement.

[6] The base year for the consumer price index is 2021, as obtained from StatsSA.

the IRP5, IT3(a), and ITR12 tax certificates while ensuring that labour income for one individual is not counted twice.

A more detailed explanation regarding the data cleaning and variable creation is available in the Appendix.

3.3 Descriptive statistics for the formal economy

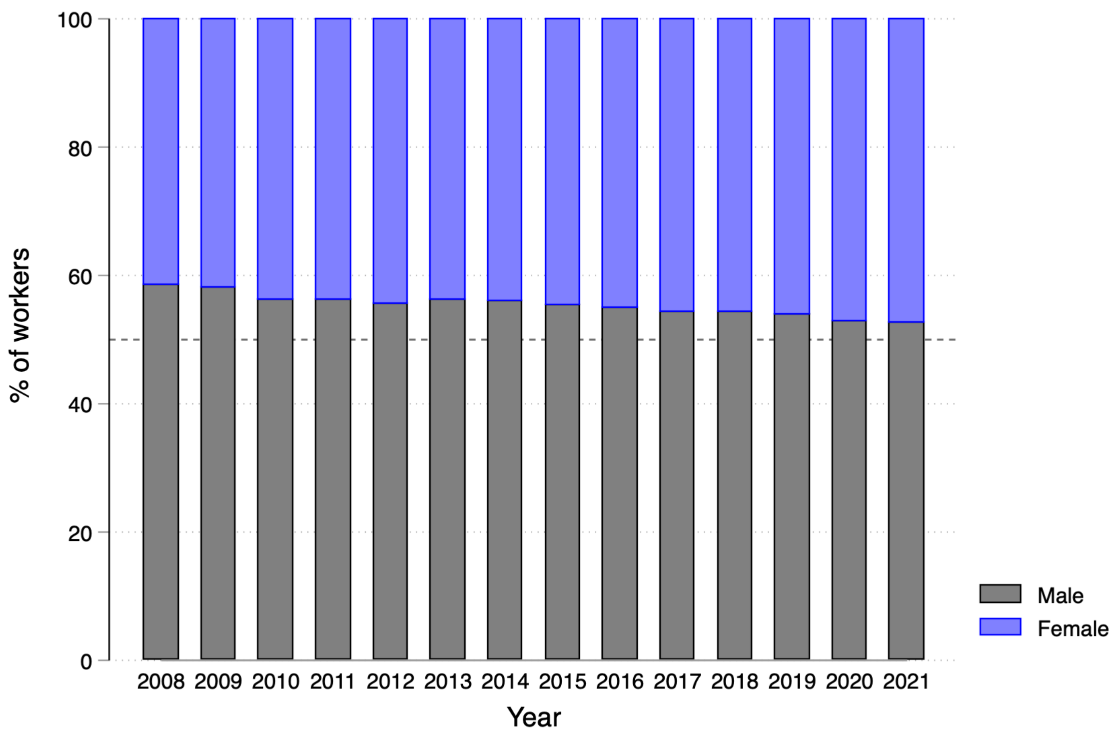
Employment gap by gender

The gender employment gap from 2008 to 2021 has been reduced, as seen in Figure 1. Even though males continue to dominate the formal labour market, there has been a consistent increase in the employment of females in the formal sector. During this time, the number of females employed in the formal sector increased by 5.96 percentage points from 41.43% in 2008 to 47.39% in 2021.

The demand side of the labour market shows that formal employment numbers are equalizing over time (Figure 1). To determine whether the gender gaps are equalizing on the supply side of the labour market, we consider the labour force participation and unemployment rates by gender.

We note that the number of women participating in the labour force has had a positive trend since 2008, as seen in Table C6 (ILO 2022). We hypothesize that females might be more active in the labour force due to households requiring more than one income to meet expenses, women divorcees needing income to support their families after being separated from a spouse, and a change in social norms, i.e. societal norms frowning upon women who work are slowly diminishing (World Economic Forum 2022).

Figure 1: Per cent of employed people by gender (2008–21)



Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

The labour force participation rate[7] for males was 69.97% in 2008, and the female labour force participation in 2008 was 56.03% (ILO 2022).[8] In 2021, the labour force participation for males was 65.25%, 11.21 percentage points higher than the female labour force participation (54.04%). Further, the labour force participation gender gap between males and females with basic education was 13.31 percentage points in 2021 and is larger relative to those with higher education (7.29 percentage points). This education gap by gender may be one driver of women having a smaller presence in the labour force relative to men.

The unemployment rates for both men and women had an increasing trend from 2008 to 2021. However, the unemployment gap by gender decreased over this period. The female unemployment rate was 4.36 percentage points higher than the male unemployment rate in 2008, but this gap dropped to 3.06% in 2021 (Table C6).

Taking these findings together, women face more challenges to enter formal employment relative to men. Women do not enter the labour force at the same rate as men and are less likely to find employment once actively searching for employment.

Median income differential by gender

Figure 2 shows that, during the 2008–20 period, the median labour income had an upward trend, indicating that employees' real incomes increased in the formal economy until 2020, whereafter real incomes (prices adjusted to 2021) decreased due to the COVID-19 pandemic. In 2008, the male median income was ZAR44,121.12, and the female median income was ZAR39,268.32 (women earned 89 cents for every ZAR1 that men earned). Over the period shown, men continued to earn more income than women, and the median income difference by gender increased over this period. In 2021, the male median income was ZAR88,989.12, and the female median income was ZAR69,480.77 (women received 78 cents relative to the ZAR1 that a man earned).

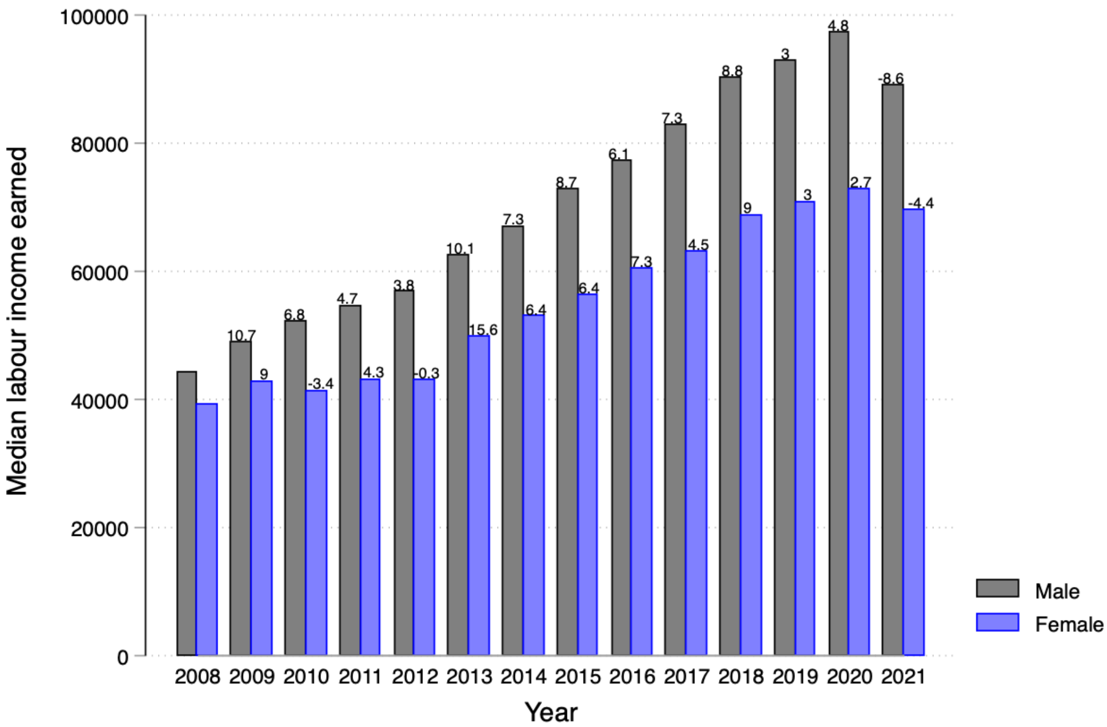
The diverging median income by gender is due to the median income growth rates: in 8 of 13 years, men had higher median income growth rates. This, together with the higher base value of male median income relative to female median income in 2008, increases the gender pay gap.

From Figure 1 and Figure 2, we note that the number of women in formal employment increases, and over the same period, the gender wage differential increases.

[7] The labour force participation rate is the proportion of individuals between the ages of 15 and 64 that is active in the labour force, by gender.

[8] These figures are available in the Appendix in Table C6.

Figure 2: Median labour income earned by gender (2008–21)



Note: the values above the bars indicate the year-on-year growth rate of median income for every year by gender.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

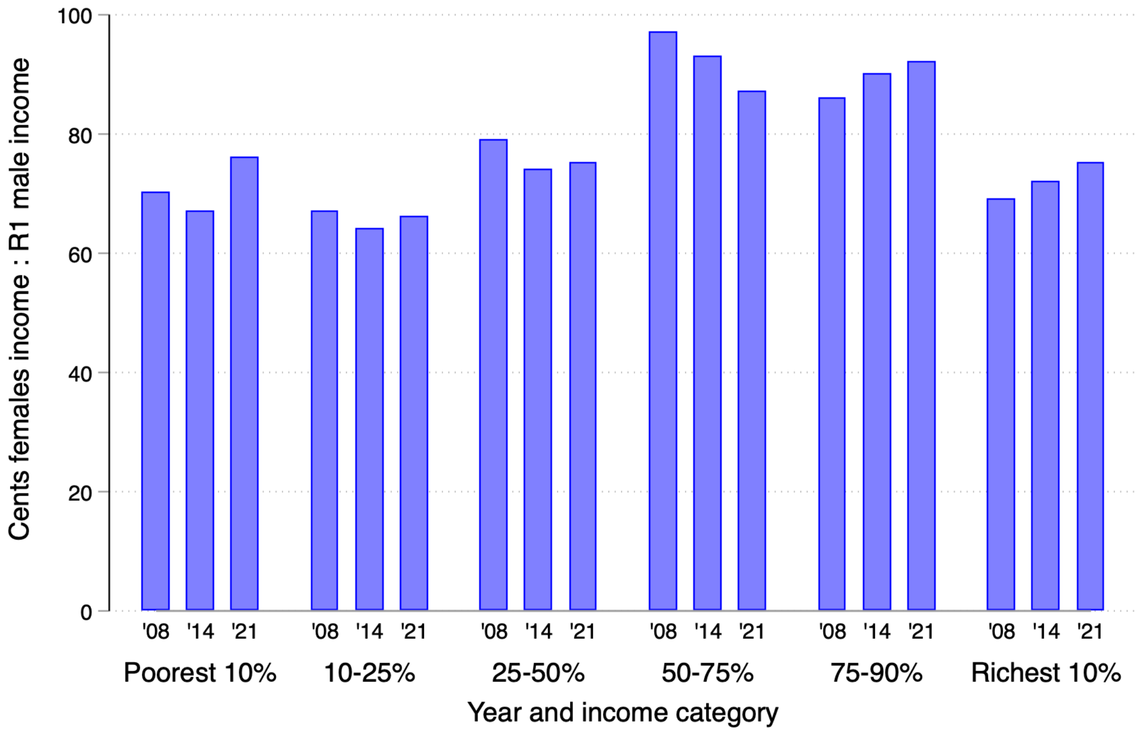
Gender pay gap by income categories

Figure 3 displays the distribution of formally employed individuals by gender and income category, highlighting the areas within the income distribution where the largest wage differential occurs. The figure is organized into six income categories: 'poorest 10%', '10–25%', '25–50%', '50–75%', '75–90%', and 'richest 10%'. For the years 2008, 2014, and 2021, we calculate the cents that women earn on average for every ZAR1 of male income earned in each income category.

Based on Figure 3, it is evident that the gender pay gap persists across all income categories in 2021. Furthermore, the trends in these income categories vary, showing the importance of observing the wage differential across the distribution.

On average, the tails (i.e. 'poorest 10%', '10–25%', and 'richest 10%') of the income distribution are the most unequal relative to the middle of the distribution. The progress towards income parity varies across different income categories. Specifically, the '75–90%' and 'richest 10%' income categories consistently showed movement towards equalizing the gender pay gap from 2008 to 2021.

Figure 3: Female cents earned per ZAR1 male income by gender and income category (2008–21)



Note: we organize the population into male and female subset populations. Then, we divide males and females into six income categories to determine where male and female workers are placed in the income distribution. We divide the female median income by the male median income for every income category to obtain the ratio of female income to ZAR1 of male labour income earned.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

In the 'poorest 10%' income category, females experienced a notable increase in earnings relative to male labour income from 2014 to 2021. Specifically, they earned an additional 9 cents per ZAR1 of male labour income during this period. This increase may be attributed to the implementation of the National Minimum Wage Act, which came into effect on 1 January 2019 and may have contributed to reducing income inequality within this income category (Republic of South Africa 2018).

It is worth noting that the increase in earnings for females in the 'poorest 10%' category from 2014 to 2021 represents the highest increase among all income categories during that period.[9]

High-skill population

In the formal economy, we note that the tails of the income distribution have a high wage discrepancy by gender relative to the middle of the distribution. It is likely that we cannot account for all income in the lower tail of the income distribution as we do not have access to the informal sector. Therefore, we focus our further investigation on the upper tail of the distribution by looking at the high-skill population specifically. Mattos and Ogura (2009) show that the majority of income earned in the high-skill sector is accounted for in the Brazilian formal economy.

In the IRP5/IT3(a) data set, individuals have a four-digit Standard Industrial Classification (SIC) code corresponding to the primary economic activity of the employing firm. The data set spans from 2008

[9] In the Appendix, Table C8 contains the median income and growth rates for all income categories.

until 2021; however, the industry code variable is sparsely populated before 2015.[10] Therefore, we can reliably determine the industry for which a labourer works from 2015 onwards to identify high-skill workers. We use the definition by Chen et al. (2022) to recognize which industries predominantly represent high skills. The authors determine high-skill industries in the American economy where sectors are categorized with the North American Industry Classification System (NAICS) codes. The categories of the SIC codes correspond well with the NAICS codes.

If individuals submit one tax certificate, then we can identify whether the tax certificate belongs to a high-skill industry. When an individual has multiple tax certificates, we check whether each certificate belongs to the same industry. Individuals with multiple certificates that relate to the same industry can be classified as working in a high-skill industry or not. However, individuals with multiple certificates that have industry codes aligning with both the high-skill industry and other industries are removed from this analysis.[11]

High-skill definition. Chen et al. (2022) identify 10 industries that require the highest labour skills in the United States. These industries are: 1) 'Legal Services'; 2) 'Research, Development, and Testing Services'; 3) 'Engineering, Architectural, and Surveying'; 4) 'Educational Services'; 5) 'Offices and Clinics of Doctors of Medicine'; 6) 'Accounting, Auditing, and Bookkeeping Services'; 7) 'Management and Public Relation Services'; 8) Health Services'; 9) 'Mailing, Reproduction, Commercial Art and Photography, and Stenographic Services'; and 10) 'Advertising'. We use these industries to identify whether individuals work in a high-skilled sector.

An individual is coded as working in a high-skill industry in South Africa if their tax certificate belongs to one of the following SIC codes: 1) 'Legal Activities' (691); 2) 'Scientific Research and Development' (72); 3) 'Architectural and Engineering activities, Technical Testing and Analysis' (71); 4) 'Education Services' (85); 5) 'Human Health Activities' (86); 6) 'Accounting, Bookkeeping, Auditing Activities, and Tax Consultancy' (692); 7) 'Activities of Head Offices and Management Consultancy Activities' (70); 8) 'Commercial Art and Photography' (74); or 9) 'Advertising and Market Research' (73).[12] The SIC code associated with each tax certificate indicates the industry that the firm operates in.

Table 1 presents the number of workers for each sector in 2015 and 2021. The table also displays the median income of each industry. The education industry is the largest employer of high skills, followed by the health care industry over this period. The 'Architectural and Engineering activities' sector and the 'Scientific Research and Development' industry have the highest median incomes for 2015 and 2021.

As South Africa is a developing country, we had concerns about the status of the 'Commercial Art and Photography' sector relative to other industries categorized as high-skill sectors. In 2015, the 'Commercial Art and Photography' sector had the smallest number of skilled workers and accounted for only 2.87% of all workers in the high-skill industry. Furthermore, the median income of high-skill individuals in this sector converged with the median income of other high-skill industries in 2021. Therefore, considering the small number of individuals within this sector and the converging median income, we continue to include the 'Commercial Art and Photography' industry in our high-skill analysis.

[10] Table A2 displays the per cent of entries that have a missing SIC code in the IRP5/IT3(a) data set.

[11] This is described in more detail in the Appendix section 'Data cleaning process'.

[12] We only incorporate nine industries in our high-skilled sector: Chen et al. (2022) include the 'Offices and Clinics of Doctors of Medicine' and 'Health Services', where the South African SIC codes have one sector ('Human Health Activities') that incorporates both of those industries. The values in the brackets indicate the SIC codes.

Table 1: Number of workers and median income by a high-skill sector

Industry	2015 N	2015 Income	2021 N	2021 Income
Legal activities	31,365	144,262.3	44,084	178,711.0
Scientific research and development	20,337	202,220.3	19,197	368,752.5
Architectural and engineering activities	45,406	204,528.5	50,109	236,673.1
Education service	196,819	218,478.7	280,348	233,245.4
Human health activities	100,696	141,109.4	144,350	182,671.0
Accounting, bookkeeping, auditing activities	39,345	162,609.8	52,680	210,214.4
Management consultancy activities	20,835	203,885.3	27,407	208,340.3
Advertising and market research	19,101	144,901.2	21,888	216,748.8
Commercial art and photography	14,009	140,640.9	20,771	183,707.0

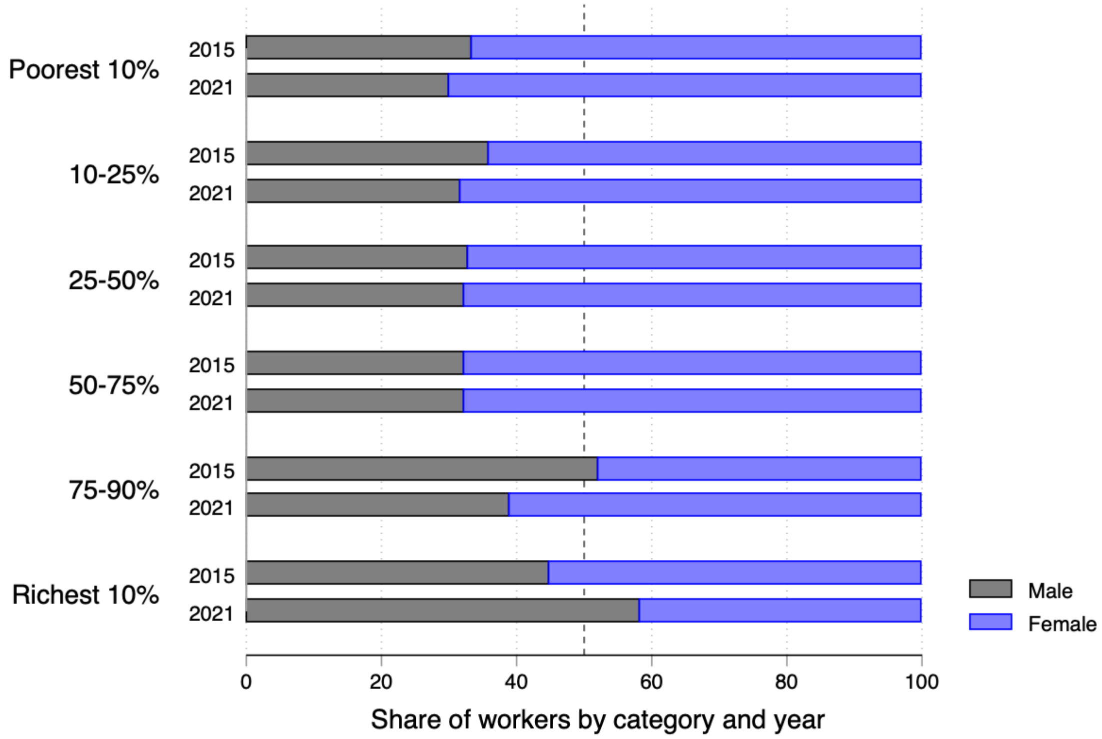
Note: this table displays the number of people in the industry (N) and the median income (Income) for every high-skill industry. The industry codes to identify high-skill sectors are poorly populated before 2015; therefore, we exclude years prior to 2015.

Source: authors' compilation using the SARS tax data (National Treasury and UNU-WIDER 2022).

High-skilled industry: employment gap by gender. Figure 4 shows that more women than men are employed in five out of the six income categories for the high-skill sector. Chen et al. (2022) include the 'Legal and Accounting Activities', 'Human Health Activities', and 'Education' sectors in the high-skill industry definition. In 2015, we find that five of the high-skill industries have a higher share of females (displayed in parentheses): 1) 'Human Health Activities' (78%), 2) 'Legal and Accounting Activities' (68%), 3) 'Education' (66%), 4) 'Advertising and Market Research' (58%), and 5) 'Scientific Research and Development' (56%).

Despite these sectors being female-dominated, the majority of females occupy lower-paying positions, as a larger share of women relative to men is evident in the low-income-earning categories. It seems that the top end of the income distribution is more variable, as women have a larger share in the 'richest 10%' income category for 2015, while in 2021, men occupy more positions in the 'richest 10%' income category.

Figure 4: Per cent of high-skill people by gender and income category for 2015 and 2021



Note: the high-skill population is organized into six income categories to determine where male and female workers are placed in the income distribution. The base sums the male and female share for each income category, which totals 100.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

The organization of men and women into the income categories can be attributed to occupational segregation. For instance, according to the World Health Organization's Global Health Workforce Statistics (2022), South Africa has 1.3 nurses per 1,000 individuals, with 89.44% of nurses being female (South African Nursing Council 2021). On the other hand, there are 0.8 physicians per 1,000 people, and physicians are primarily male. In 2019, there were 0.47 male doctors per 1,000 individuals compared to 0.32 female doctors (Tiwari et al. 2021). This occupational segregation has wage implications, as doctors tend to earn more income compared to nurses.

Similarly, 79% of primary school teachers are female (UNESCO 2023), while only 24.5% of professors in tertiary education are female (van Wyk 2018). On average, university professors earn more than primary school teachers. These figures support the idea that men tend to sort themselves disproportionately into higher income categories.

4 Empirical strategy

4.1 Linear regression: gender differential at the mean

Using a linear regression accounting for fixed effects, we estimate the relationship between gender and real income:

$$\begin{aligned} ln(real\ income)_{i,t} &= \alpha_0 + \beta_1 male_i + \beta_2 high\ skill_{i,t} + \beta_3 full\ time_{i,t} \\ &+ \beta_4 male_i \cdot high\ skill_{i,t} + \gamma \mathbf{X_{i,t}} + {}_{i,t} \end{aligned} \tag{1}$$

The dependent variable refers to the natural logarithm of real income. Male individuals are coded one in the *male* variable and zero otherwise. If an individual works in a high-skill industry,[13] then *high skill* returns one and zero otherwise. The *full-time* variable returns one when an individual works full time in one company or in one industry and zero otherwise. The individual control variables ($\mathbf{X_{i,t}}$) include the age of the individual, quadratic function of age, province that the individual lives in, industry that an individual works in, and year.

We use an estimator for linear models proposed by Correia (2016), which accounts for multi-way fixed effects. This technique incorporates fixed effects at multiple levels, including the year, industry in which an individual works, and province in which an individual lives, to address unobserved heterogeneity. We use fixed effects for the year, industry, and province to control for the role of these factors in the gender wage gap. We expect that real incomes will differ by year, province, and industry. For example, in 2021, we expect that real incomes are lower relative to the previous year's income due to COVID-19; we expect high-skill industries to have higher incomes relative to low-skill industries; and provinces that contain economic hubs should experience higher incomes relative to provinces that have high proportions of agricultural land.

4.2 Quantile regression: pay gap along the conditional income distribution

Given the variation in gender inequality across the income distribution, we use a quantile regression to better understand this variation. We estimate the relationship between gender and the natural logarithm of real income across the conditional income distribution using a quantile regression that accounts for individual effects.

$$\begin{aligned} q_\tau(\mathit{ln(real\ income)}_{i,t}) = \alpha_\tau + \beta_{1,\tau}\mathit{male}_i + \beta_{2,\tau}\mathit{high\ skill}_{i,t} + \beta_{3,\tau}\mathit{full\ time}_{i,t} \\ + \beta_{4,\tau}\mathit{male}_i \cdot \mathit{high\ skill}_{i,t} + \gamma_\tau \mathbf{X_{i,t}} + {}_{i,t,\tau} \end{aligned} \quad (2)$$

The natural logarithm of real income is the dependent variable for individual i at year t in quantile τ. The *male* variable is coded one when the individual is male and zero elsewhere. The *high-skill* variable returns one when an individual works in a high-skill industry and zero otherwise, which varies over time. Individuals can change jobs across years, where individuals can work in a high-skill industry in year t but move out of the high-skill sector in year $t{+}1$. The control variables ($\mathbf{X_{i,t}}$) include the age of the individual, quadratic function of age, year, industry in which an individual works, and province of residence.

We estimate conditional quantiles using the method of moments proposed by Machado and Silva (2019). Due to the large size of this panel data set, other approaches have difficulty estimating the quantile regression, while the method of moments is able to obtain these quantile coefficients. Further, this method is useful as it estimates the relationship of the independent variables over the entire conditional distribution.

[13] 1) 'Legal Activities'; 2) 'Scientific Research and Development'; 3) 'Architectural and Engineering Activities, Technical Testing and Analysis'; 4) 'Education Services'; 5) 'Human Health Activities'; 6) 'Accounting, Bookkeeping, Auditing Activities, and Tax Consultancy'; 7) 'Activities of Head Offices and Management Consultancy Activities'; 8) 'Commercial Art and Photography'; 9) 'Advertising and Market Research'.

5 Regression results

5.1 Gender wage differential at the mean

In Table 2, we use a linear regression to estimate the relationship between gender and the natural logarithm of real income for the entire sample of individuals from 2015 to 2021.[14]

Column 1 reveals the disparity in earnings between men and women in the formal economy, with men earning 35.66% more than women at the mean of the real income distribution. This coefficient retains its significance and robustness even after including control variables. We expect this relationship, as Figure 3 shows the sustained disparity between income by gender. This finding is in line with prior research, showing a gender wage gap in South Africa (Bhorat and Goga 2012; Mosomi 2019; Ntuli 2014).

Individuals employed in high-skill industries unsurprisingly exhibit higher levels of real income compared to those working in other sectors, as shown in column 2. Generally, high-skill sectors require tertiary education to obtain specialized knowledge to perform industry-specific tasks. Moreover, the limited availability of individuals with industry-specific skills required in high-skill sectors increases the income premium of individuals.

Column 3 takes into account an individual's age. We observe that as individuals grow older, their real income is likely to increase due to accumulated experience and greater prospects for promotion, resulting in a positive income return. The quadratic function of age demonstrates that real income growth accelerates until a certain age threshold; after this age threshold is reached, income growth declines until retirement. These findings are in line with the literature (Becker 1965).

Full-time workers earn a higher real income relative to part-time workers, as column 3 illustrates. Full-time workers receive more income instalments throughout the year compared to their part-time peers working seasonally. In South Africa, part-time workers are predominantly individuals with low levels of education working in low-skilled sectors (Leibbrandt et al. 2010) where these individuals are already at the lower end of the income distribution.

[14] The data prior to 2015 lack adequate representation of industry codes; therefore, we confine our analysis to 2015 onwards. Table A2 in the Appendix shows the number of missing SIC codes by year.

Table 2: Linear regression estimates: the impact of gender on real wage earned

	Dependent variable: natural log of real income							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male	0.305***	0.329***	0.293***	0.298***	0.248***	0.301***	0.305***	0.257***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
High skill		0.434***	0.415***	0.410***		0.451***	0.447***	
		(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	
Age			0.181***	0.179***	0.134***	0.180***	0.179***	0.134***
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Age^2			-0.002***	-0.002***	-0.001***	-0.002***	-0.002***	-0.001***
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Full time			1.781***	1.781***	1.577***	1.781***	1.781***	1.577***
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Male · High skill						-0.093***	-0.094***	-0.117***
						(0.00)	(0.00)	(0.00)
Predictive margins: real wage difference by gender and skill level								
Male						0.292***	0.296***	
						(0.00)	(0.00)	
Male · Other skill						0.301***	0.305***	
						(0.00)	(0.00)	
Male · High skill						0.208***	0.210***	
						(0.00)	(0.00)	
Province included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year included	No	No	No	Yes	Yes	No	Yes	Yes
Industry included	No	No	No	No	Yes	No	No	Yes
Obs	60,776,758	60,776,758	60,776,758	60,776,758	60,776,681	60,776,758	60,776,758	60,776,681
Adj. R^2	0.034	0.039	0.403	0.408	0.503	0.403	0.408	0.503

Note: standard errors clustered at the individual level are in parentheses. *Male* is coded one for male individuals and zero otherwise. *High skill* refers to the industries that are considered to require a high-skill level to work in. As defined by Chen et al. (2022), these industries are: 1) 'Legal Activities'; 2) 'Scientific Research and Development'; 3) 'Architectural and Engineering Activities, Technical Testing and Analysis'; 4) 'Education Services'; 5) 'Human Health Activities'; 6) 'Accounting, Bookkeeping, Auditing Activities, and Tax Consultancy'; 7) 'Activities of Head Offices and Management Consultancy Activities'; 8) 'Commercial Art and Photography'; and 9) 'Advertising and Market Research'. *Age* indicates the age of the individual, and Age^2 is the quadratic function of the variable *age*. *Full time* returns one when the individual is a full-time employee and zero otherwise. *Province* indicates where the individual lives. *Year* controls for the year. *Industry* controls for the industry that an individual works in using a five-digit SIC code. * p<0.10, ** p<0.05, *** p<0.01

Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Columns 4 and 5 control for the year and industry, respectively, while maintaining the robustness of all previous relationships. Recall that the high-skill variable returns one for industries identified as high-skilled from the industry codes and zero elsewhere. When including the industry as a control variable, the high-skill variable is omitted since it is linearly related to the industry variable.

The interaction between gender and working in a high-skill sector is introduced in column 6. The income premium for being male and working in a high-skill sector is 0.208, indicating that men earn 23.12% more compared to women in high-skill sectors. This finding likely reflects an occupational segregation gap, where women may choose occupations in high-skill sectors that offer lower incomes relative to careers preferred by men, which are typically higher paying. This relates to Figure 4, where the majority of women occupy the lower-income categories and men tend to organize themselves into the higher-earning-income categories. For instance, in South Africa, the share of workers in the medical industry are split so that the majority of women work as nurses, while men are predominantly doctors (Tiwari et al. 2021). Bhorat and Goga (2013) find that the gender wage differential was smaller at the 70th, 80th, and 90th quantiles relative to the lower tail of the income distribution due to higher education and professional work. This is in line with our findings.

In contrast, men working in the ‘other-skill’ sector earn 35.12% more than women in the ‘other-skill’ sector, as seen in column 6. Therefore, the gender pay gap is smaller in the high-skill sector compared to the other-skill industry. We hypothesize that the requirement for higher education in the high-skill sector may contribute to the smaller gender pay differential. The gender income differential may speak to the gender difference in managerial positions. The percentage of women in managerial positions was 31.6% in 2021, which increased by 2 percentage points from 2008 (StatsSA 2022). As men had more executive positions relative to women over this period, this may be a contributing factor to the persistent gender pay gap.

Columns 7 and 8 include the year and industry control variables, respectively, while preserving the robustness of all relationships.

5.2 Gender pay gap along the conditional income distribution

As gender inequality varies over the income distribution, we use a quantile regression to explore this variation. Figure 5 displays the coefficient estimates of each explanatory variable for the conditional quantiles of the natural logarithm of real income.[15]

The male coefficient is positive and statistically significant for every conditional quantile when controlling for the individual’s age, quadratic function of age, full-time employment status, year, industry, and province. That is, men earn higher incomes compared to women across the conditional real income distribution. This is expected, as women only earn a portion of men’s income in every income category, as seen in Figure 3.

The wage discrepancy is highest at the first quantile (0–10) of the conditional income distribution. We expect this result because females predominantly submit IT3(a) tax certificates over IRP5 certificates, while males are more likely to submit IRP5 certificates over IT3(a) tax certificates.[16] The IT3(a) certificate is associated with lower incomes. In line with our findings, studies found that the left tail of the conditional income distribution has the largest wage discrepancy (Bhorat and Goga 2013; Mosomi 2019). The first quantile of the conditional wage distribution shows a gender wage gap of 28.15%. We hypothesize that this may be a lower-bound estimate, as these data do not capture the informal economy. In the informal economy, it is estimated that males earn 33% more income than females (StatsSA 2019).

[15] The corresponding quantile regression table is available in the Appendix in Table C9.

[16] Table B4 displays these figures.

Furthermore, men constitute the majority of individuals working in the informal sector, accounting for 56.24% (StatsSA 2019). In this sector, men are more likely to start a business (73.9%), while women are more willing to help in the family business without pay (63.0%) (StatsSA 2019).

Furthermore, females may choose to work in industries that pay wages comparable to the minimum wage, whereas males are more likely to work in sectors that offer slightly higher wages than the minimum wage (ZAR23.19 in 2022) (South African Government 2022). For instance, the retail sector (industry code 47) is dominated by female workers, with 59% of all workers in the retail industry being female in 2018. The average pay for a low-earning worker in the retail sector, such as a cashier, was ZAR23.45 per hour in 2022 (Payscale 2023b). On the other hand, the construction industry (industry code 41) has 83% male workers, and the average hourly income for a bricklayer was ZAR31.73 in 2022 (Payscale 2023a).

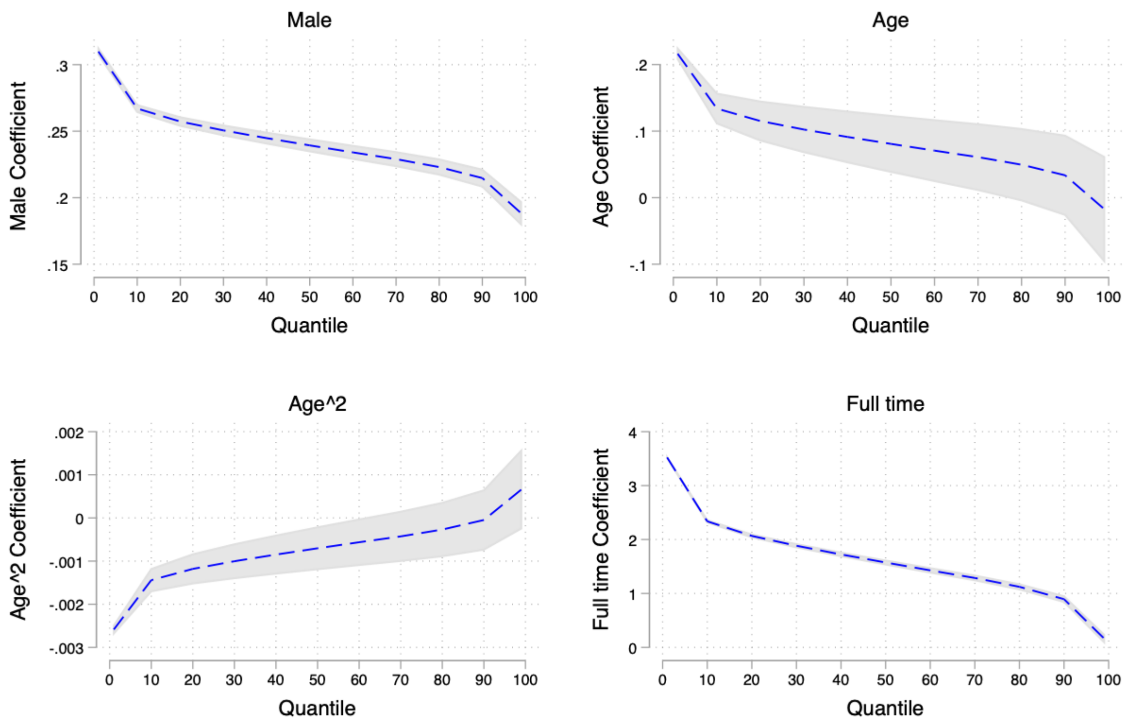
From 2008 to 2021, the wage differential by gender consistently decreased in the '75–90%' and 'richest 10%' income categories, as observed in Figure 3. This reduction can be a contributing factor to the smaller male coefficient across the right tail of the conditional income distribution.

The relationship of age with the conditional real income distribution follows a common result from the literature for the conditional income distribution to the left of the 90$^{\text{th}}$ percentile. Age has a positive relationship with real income across the conditional income distribution except for the 'richest 10%' income category. As individuals age, they are likely to increase their earnings due to experience (Becker 1965). The quadratic age function has a negative relationship with real income, illustrating that income growth will slow down at some point in the working life, and eventually the income growth will be negative close to retirement age for the conditional real income distribution to the left of the 'richest 10%' income category.

Specific to the '90–100%' (richest 10%) conditional quantile, the quadratic function of age is positive, showing that real income continues to grow as an individual ages. We hypothesize that, as individuals age, they become more experienced and are more likely to occupy C-suite positions that are correlated with higher income and where high annual bonuses are paid out.

In the '0–10%' conditional quantile of the distribution, an individual with a full-time job has a higher real income relative to a part-time worker. In agriculture, seasonal workers may not find employment during the offseason, limiting their income to the harvesting season, while full-time workers earn an income throughout the year. In contrast, the '90–100%' conditional quantile shows a minor premium for full-time work on real income. This conditional quantile may include individuals who work as industry experts on a part-time basis for multiple companies (such as directors), where the income can be similar to the real income earned from full-time employment.

Figure 5: Quantile regressions: influence of factors over the conditional income distribution



Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

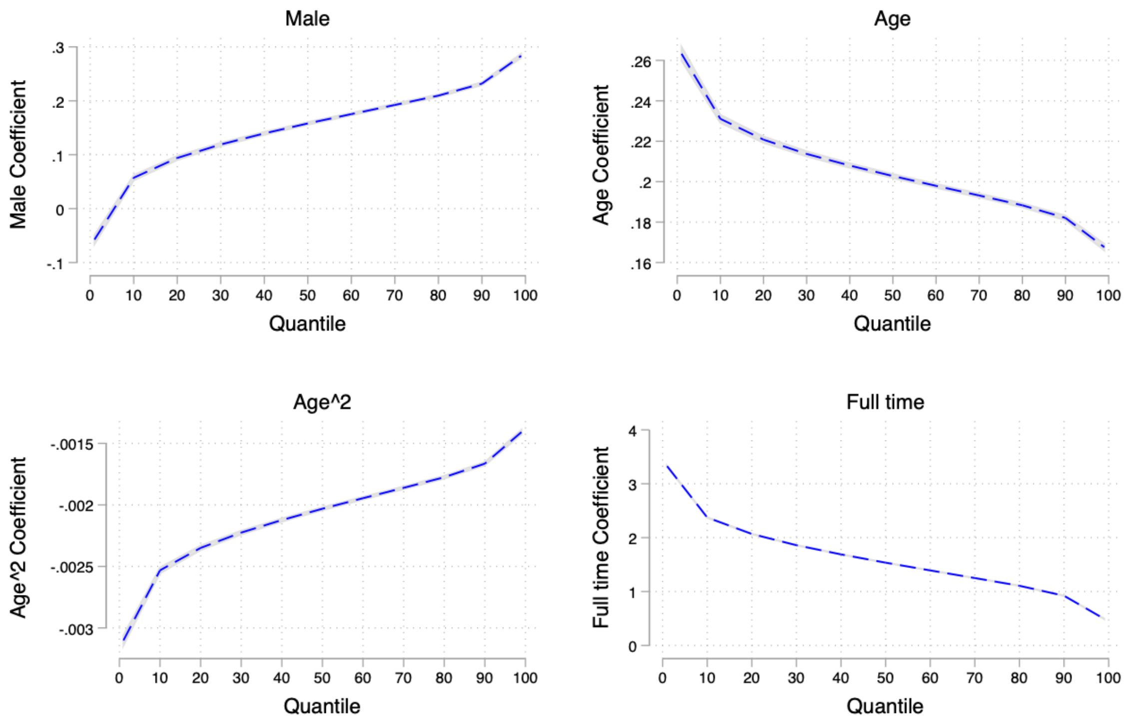
5.3 High-skill population: income differential along the conditional income distribution

Figure 6 replicates the analysis in Section 4.2 for the high-skill sector specifically. This figure shows the coefficient estimates of each independent variable for the conditional quantiles of the natural logarithm of real income in the high-skill sector only.[17]

The high-skill sector is identified from the industry codes that the firm operates in. However, the firm can hire employees that perform general activities that are not specific skills for the industry. For example, firms employ cleaning staff, but these skills are not specific to the industry.

[17] The corresponding quantile regression table is available in the Appendix in Table C10. Recall that the high-skill sector includes: 1) 'Legal Activities'; 2) 'Scientific Research and Development'; 3) 'Architectural and Engineering Activities, Technical Testing and Analysis'; 4) 'Education Services'; 5) 'Human Health Activities'; 6) 'Accounting, Bookkeeping, Auditing Activities, and Tax Consultancy'; 7) 'Activities of Head Offices and Management Consultancy Activities'; 8) 'Commercial Art and Photography'; and 9) 'Advertising and Market Research'.

Figure 6: Quantile regressions: influence of factors over the conditional income distribution for the high-skill population



Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

The male coefficient shows the premium that men earn across the conditional real income distribution. The range of the male coefficient in the first (0–10%) conditional quantile contains negative and positive values, showing the variability of this conditional quantile. At the left tail of the conditional income distribution, the salaries for support staff, trainees, and junior staff are more equal by gender relative to the right tail of the conditional income distribution.

By observing the male coefficient when the conditional real income distribution is above the median, we improve the likelihood of removing general workers from the sample and focusing on individuals involved in industry-specific tasks where these individuals would require tertiary education. The conditional real income distribution at the median displays that men earn 17.11% more real income relative to women.[18] This pay disparity becomes larger, moving towards the right tail (90–100%) of the conditional real income distribution, where men earn 26.11% more than women.[19] We note that men fill the majority of management positions (Labour Force Survey 2021), and in a recent study, females in top executive positions earn 72 cents for every ZAR1 of male income (Price Waterhouse Coopers 2020).

The coefficients for *age*, *age*^{2}, and *full time* exhibit similar patterns when compared to the quantile regressions for the full sample (Figure 5). Therefore, the same interpretations as those from Figure 5 apply to these variables in the high-skill sample, too.

[18] The male coefficient from the quantile regression at the 50th conditional quantile is 0.158.

[19] The male coefficient from the quantile regression in Table 6 is 0.232 for the 90th conditional quantile.

6 Conclusion

Our paper has examined the gender wage gap in the South African formal economy and explored whether women are making progress towards financial independence over time. In contrast to previous studies that have relied on survey data, we have used administrative data, which provide a comprehensive representation of workers in the formal sector. It is worth noting that this type of analysis is relatively scarce in developing countries. A search retrieves only one other paper using administrative data to observe the level of inequality in Uruguay.

Our findings indicate that, although the number of women in the labour force has increased, there is still work to be done to achieve gender parity. The share of women in the labour force has grown over the years, but the wage disparity has also widened. Men continue to earn higher median incomes and experience faster income growth compared to women. In 2008, women earned approximately 89% of men's wages, but by 2021, this ratio had dropped to 78%. Additionally, our analysis of the high-skill sector reveals that women are underrepresented in the top income category of the income distribution in 2021.

At the mean of the income distribution, men consistently earn notably higher incomes than women. However, we find that the gender pay gap is smaller in the high-skill sector compared to other sectors of the economy. We hypothesize that individuals need tertiary education to enter this high-skill sector, where it is likely that education levels are more similar by gender relative to other industries in the formal economy.

Analysing the conditional quantiles across the income distribution, we observe that the income disparity is most pronounced at the 10^{th} percentile, suggesting that women may be more prevalent in industries with wages closer to the minimum wage, while men tend to self-select into sectors with higher wage rates.

Furthermore, when focusing specifically on the high-skill sector, we find that the largest income difference by gender occurs at the top end of the conditional income distribution, potentially reflecting the low representation of women in managerial positions within this sector.

While administrative data provide valuable insights by including all formal sector workers in estimating the gender wage differential, it does have limitations. Notably, it lacks information on factors such as education and experience, restricting us from assessing the impact of these human capital variables on the gender wage gap. To gain a more comprehensive understanding of the wage differential in South Africa, future research should also include the informal economy, where women's progress in achieving financial independence may differ from that in the formal sector.

References

Anderson, S., and Eswaran, M. (2009). 'What Determines Female Autonomy? Evidence from Bangladesh'. *Journal of Development Economics*, 90(2): 179–91. https://doi.org/10.1016/j.jdeveco.2008.10.004

Bargain, O., Etienne, A., and Melly, B. (2018). 'Public Sector Wage Gaps Over the Long-Run: Evidence from Panel Administrative Data'. IZA Discussion Paper 11924. Bonn, Germany: IZA. https://doi.org/10.2139/ssrn.3286184

Bayard, K., Hellerstein, J. K., and Troske, K. R. (2004). 'New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employer-Employee Data'. *Sex Differences in Labor Markets*, 21(4): 299–332. https://doi.org/10.1086/377026

Becker, G. S. (1965). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: National Bureau of Economic Research.

Bhorat, H., and Goga, S. (2012). 'The Gender Wage Gap in the Post-Apartheid South African Labour Market'. DPRU Working Paper 12/148. Cape Town: Development Policy Research Unit, University of Cape Town. https://doi.org/10.2139/ssrn.2184258

Bhorat, H., and Goga, S. (2013). 'The Gender Wage Gap in Post-Apartheid South Africa: A Re-Examination'. *Journal of African Economies*, 22(5): 827–48. https://doi.org/10.1093/jae/ejt008

Blau, F. D., and Kahn, L. M. (2017). 'The Gender Wage Gap: Extent, Trends, & Explanations'. *Journal of Economic Literature*, 55(3): 789–865. https://doi.org/10.1257/jel.20160995

Burdín, G., De Rosa, M., Vigorito, A., and Vilá, J. (2022). 'Falling Inequality and the Growing Capital Income Share: Reconciling Divergent Trends in Survey and Tax Data'. *World Development*, 152(-): . https://doi.org/10.1016/j.worlddev.2021.105783

Casale, D., and Posel, D. (2002). 'The Continued Feminisation of the Labour Force in South Africa: An Analysis of Recent Data and Trends'. *South African Journal of Economics*, 70(1): 156–84. https://doi.org/10.1111/j.1813-6982.2002.tb00042.x

Chen, J. Z., Jang, Y., Jung, B., and Noh, M. (2022). 'Labor Skill and Accounting Conservatism'. Available at SSRN. https://doi.org/10.2139/ssrn.4017970

Correia, S. (2016). 'reghdfe: Estimating Linear Models with Multi-Way Fixed Effects'. In *2016 Stata Conference 24*.

Danquah, M., Iddrisu, A. M., Boakye, E. O., and Owusu, S. (2021). 'Do Gender Wage Differences Within Households Influence Women's Empowerment and Welfare? Evidence from Ghana'. *Journal of Economic Behavior and Organization*, 188(-): 916–32. https://doi.org/10.35188/UNU-WIDER/2021/978-5

Donald, A., Koolwal, G., Annan, J., Falb, K., and Goldstein, M. (2020). 'Measuring Women's Agency'. *Feminist Economics*, 26(3): 200–26. https://doi.org/10.1080/13545701.2019.1683757

Engels, F. (1972). *The Origin of the Family, Private Property, and the State*. New York: Pathfinder Press (original work published 1884).

Eswaran, M. (2002). 'The Empowerment of Women, Fertility, and Child Mortality: Towards a Theoretical Analysis'. *Journal of Population Economics*, 15(-): 433–54. https://doi.org/10.1007/s001480100070

Fortin, N. M., Bell, B., and Böhm, M. (2017). 'Top Earnings Inequality and the Gender Pay Gap: Canada, Sweden, and the United Kingdom'. *Labour Economics*, 47(February): 107–23. https://doi.org/10.1016/j.labeco.2017.05.010

Fransen, E., Plantenga, J., and Vlasblom, J. D. (2012). 'Why Do Women Still Earn Less Than Men? Decomposing the Dutch Gender Pay Gap, 1996–2006'. *Applied Economics*, 44(33): 4343–54. https://doi.org/10.1080/00036846.2011.589818

Gradín, C. (2021). 'Occupational Gender Segregation in Post-Apartheid South Africa'. *Feminist Economics*, 27(3): 102–33. https://doi.org/10.1080/13545701.2021.1906439

Grün, C. (2004). 'Direct and Indirect Gender Discrimination in the South African Labour Market'. *International Journal of Manpower*, 25(3–4): 321–42. https://doi.org/10.1108/01437720410541425

Hanmer, L., and Klugman, J. (2016). 'Exploring Women's Agency and Empowerment in Developing Countries: Where Do We Stand?' *Feminist Economics*, 22(1): 237–63. https://doi.org/10.1080/13545701.2015.1091087

Hospido, L., and Moral-Benito, E. (2016). 'The Public Sector Wage Premium in Spain: Evidence from Longitudinal Administrative Data'. *Labour Economics*, 42(-): 101–22. https://doi.org/10.1016/j.labeco.2016.08.001

ILO (2022). *Labour Force Statistics Database (LFS) ILOSTAT*. International Labour Organization. Available at: https://ilostat.ilo.org/data/

Kerr, A. (2020). 'Earnings in the South African Revenue Service IRP5 Data'. WIDER Working Paper 2020/62. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/819-1

Kunze, A. (2017). 'The Gender Wage Gap in Developed Countries'. IZA Discussion Paper 10826. Bonn: IZA. https://doi.org/10.2139/ssrn.2988173

Landman, M. S., and O'Clery, N. (2020). 'The Impact of the Employment Equity Act on Female Inter-Industry Labour Mobility and the Gender Wage Gap in South Africa'. WIDER Working Paper 2020/52. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2020/809-2

Leibbrandt, M., Woolard, I., McEwen, H., and Koep, C. (2010). *Employment and Inequality Outcomes in South Africa*. Cape Town: Southern Africa Labour and Development Research Unit (SALDRU) and School of Economics, University of Cape Town.

Machado, J. A., and Silva, J. S. (2019). 'Quantiles Via Moments'. *Journal of Econometrics*, 213(1): 145–73. https://doi.org/10.1016/j.jeconom.2019.04.009

Mattos, E., and Ogura, L. M. (2009). 'Skill Differentiation Between Formal and Informal Employment'. *Journal of Economic Studies*, 36(5): 461–80. https://doi.org/10.1108/01443580910992384

Mosomi, J. (2019). 'Distributional Changes in the Gender Wage Gap in the Post-Apartheid South African Labour Market'. WIDER Working Paper 2019/17. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2019/651-7

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., and Handelsman, J. (2012). 'Science Faculty's Subtle Gender Biases Favor Male Students'. *Proceedings of the National Academy of Sciences*, 109(41): 16474–9. https://doi.org/10.1073/pnas.1211286109

Mukhwana, A. M., Abuya, T., Matanda, D., Omumbo, J., and Mabuka, J. (2020). 'Factors Which Contribute to or Inhibit Women in Science, Technology, Engineering, and Mathematics in Africa'. *The African Academy of Sciences*,(April): 1–58. Available at: https://www.aasciences.africa/

National Treasury and UNU-WIDER (2022). 'IRP5 Worker-Level Data 2008–2021 [dataset].'. Version 5.0 (beta). Pretoria: South African Revenue Service [producer of the original data], 2022. Pretoria: National Treasury and UNU-WIDER [producer and distributor of the harmonized data set], 2022.

Ntuli, M. (2014). 'Exploring Gender Wage "Discrimination" in South Africa, 1995–2004: A Quantile Regression Approach'. Tech. Rep. No. IPC Working Paper Series Number 56. Ann Arbor, MI: IPC.

Payscale. (2023a). *Average Bricklayer Hourly Pay in South Africa.* Available at: https://www.payscale.com/research/ZA/Job=Bricklayer/Hourly_Rate (Accessed on: May 19, 2023)

Payscale. (2023b). *Average Cashier Hourly Pay in South Africa.* Available at: https://www.payscale.com/research/ZA/Job=Cashier/Hourly_Rate (Accessed on: May 19, 2023)

Pedace, R. (2000). 'Using Administrative Records to Assess Earnings Reporting Error in the Survey of Income and Program Participation'. *Journal of Economic and Social Measurement*, 26(3–4): 173–92.

Price Waterhouse Coopers (2020). *Practices and Remuneration Trends Report.*

Republic of South Africa (2018). 'Act No. (of 2018: National Minimum Wage Act, 2018'. Tech. Rep.. 10.1136/inpract.26.2.110-a

Rice, L., and Barth, J. M. (2016). 'Hiring Decisions: The Effect of Evaluator Gender and Gender Stereotype Characteristics on the Evaluation of Job Applicants'. *Gender Issues*, 33(-): 1–21. https://doi.org/10.1007/s12147-015-9143-4

Roberts, G., and Schöer, V. (2021). 'Gender-Based Segregation in Education, Jobs and Earnings in South Africa'. *World Development Perspectives*, 23(June): . https://doi.org/10.1016/j.wdp.2021.100348

South African Government (2022). *Minister Thulas Nxesi Announces 2022 National Minimum Wage Increases.* Available at: https://www.gov.za/speeches/minister-thulas-nxesi-announces-2022-national-minimum-wage-increases-8-feb-2022-0000 (Accessed: 25 June 2023)

South African Nursing Council (2021). *Provincial Distribution of Nursing Manpower Versus the Population of the Republic of South Africa as at 31 December 2021.* Available at: https://www.sanc.co.za (Accessed 22 January 2023)

StatsSA (2019). *Gender Series Volume VII: Informal Economy, 2013–2019* (Vol. VII). Pretoria: Statistics South Africa. Available at: www.statssa.gov.za

StatsSA (2022). *Quarterly Labour Force Survey: Quarter 2 2022.* Pretoria: Statistics South Africa.

Tiwari, R., Wildschut-February, A., Nkonki, L., English, R., Karangwa, I., and Chikte, U. (2021). 'Reflecting on the Current Scenario and Forecasting the Future Demand for Medical Doctors in South Africa Up to 2030: Towards Equal Representation of Women'. *Human Resources for Health*, 19(1): 1–13. https://doi.org/10.1186/s12960-021-00567-2

Travis, M. A. (2015). 'Disabling the Gender Pay Gap: Lessons from the Social Model of Disability'. *Denver University Law Review*, 91(4): 893–923.

UNESCO (2023). *Percentage of Female Teachers by Teaching Level of Education.* Available at: http://data.uis.unesco.org/ (Accessed January 22, 2023)

van Wyk, A. (2018). 'How Many Professors Are There in SA?'. *Africa Check*, 18 August 2014.

World Economic Forum (2022). *Global Gender Gap Report 2022*. Available at: https://www.weforum.org/reports/global-gender-gap-report-2022

Zawaira, T., Clance, M., and Chisadza, C. (2022). 'Social Institutions, Gender Attitudes and Female Labour Force Participation in Sub-Saharan Africa'. *South African Journal of Economics*, 91(2): 186–213. https://doi.org/10.1111/saje.12336

Appendix

A Data cleaning process

The IRP5/IT3(a) data set contains all tax certificates that are submitted to SARS by an employer on behalf of an individual. Individuals can submit more than one tax certificate. For example, if an individual works for more than one entity in a given tax year, then this individual has multiple tax certificates for the one tax year. To create a panel data set of the IRP5 and IT3(a) tax certificates, we append all available years (2008–21) by the unique tax certificate number and tax year. This panel contains all tax certificates for an individual over the tax year. In this case, the same individual may appear more than once in a tax year due to multiple tax certificates.

Table A1: Number of individuals in the IRP5/IT3(a) data set by year

Year	Population	Sample	% remains
2008	14 820 719	11 638 425	78.53
2009	15 322 791	12 257 583	80.00
2010	15 061 818	13 605 000	90.33
2011	15 891 909	14 291 302	89.93
2012	16 770 838	14 989 278	89.38
2013	16 926 058	13 643 529	80.61
2014	17 531 152	13 896 427	79.27
2015	20 068 940	15 815 900	78.81
2016	19 339 741	14 870 370	76.89
2017	20 357 287	15 185 694	74.60
2018	19 632 250	14 679 094	74.77
2019	19 581 308	14 957 269	76.39
2020	19 362 854	14 971 559	77.32
2021	17 617 738	13 687 809	77.69

Note: population refers to the number of people appearing in the IRP5 data set when individuals have an identification number associated with their tax certificate. Sample shows the number of individuals that submitted a tax certificate disclosing labour income and this individual does not have a missing gender variable.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Using the IRP5/IT3(a) data set, we remove all individuals that do not have an identification number or passport number as we would not be able to identify which tax certificate belongs to which individual. Also, since IRP5 and IT3(a) certificates can be revised, we only consider the last revision per tax certificate and remove all other versions. After deleting these observations, we count the number of individuals in every year to yield the 'Population' numbers in the IRP5/IT3(a) data set as seen in Table A1. We further refine this population by excluding all individuals not earning labour income and delete individuals with a missing gender variable which produces the 'Sample' numbers in Table A1.

A1 Creation of variables

Kerr (2020) identified all income source codes that are linked to earning labour income. The full classification of which source codes are included in the labour income variable is explained in Kerr (2020). This is our labour income estimate for all individuals.

Table A2: The % of missing SIC codes in the IRP5/IT3(a) data set

Year	% of missing SIC codes
2008	100
2009	100
2010	100
2011	100
2012	100
2013	100
2014	100
2015	10.06
2016	8.52
2017	2.18
2018	2.39
2019	7.55
2020	2.19
2021	2.35

Note: SIC codes are Standardized Industry Classification codes are 5-digit codes which highlight the the primary activity that firms are engaged in.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

We are interested in determining which individuals are associated with the high-skill sector as we evaluate the pay gap specific to this sector. To determine whether an individual belongs to a high-skill sector, we use industry codes (SIC codes). In cases where industry codes are missing in the data set, we perform imputation. If an individual works for the same firm throughout the entire period, they should have the same industry code for that period. Additionally, if all employees in one year report the same industry code for their employer, individuals with missing industry codes working for the same employer should have the same industry code as the other employees of the firm for the given year. After this imputation process, we identify the number of missing SIC codes in the data set, as shown in Table A2. We also check whether all submitted tax certificates linked to one individual have the same industry code. Prior to 2015, the data set does not have industry codes, therefore all industry analysis is completed from 2015 to 2021.

We assign individuals to the high-skill sector using the high skill definition proposed by Chen et al. (2022). As individuals can have different tax certificates for different firms, we determine whether the tax certificates are for the same industry. If individuals have many tax certificates filled for the same industry, then the total labour income will correspond to the one industry for which all the tax certificates belong. If an individual has tax certificates which belong to different industries, then the total labour income earned by this individual will have a missing industry code.

Every tax certificate contains variables related to the period of work, which we use to determine whether an individual is a full-time or part-time worker. Each certificate contains two variables. The first variable, 'Total periods in the year of assessment', identifies the total time that an individual can work in the year. Companies can define their own total period, with some certificates indicating the number of days, weeks, or months that an individual can work for their entity. The second variable, 'Total periods worked in the year of assessment', identifies the period that an individual worked in this entity for one year.

Table A3: Division of individuals into full-time and part-time workers

	Classification of full-time variable			
Year	1	2	3	4
2015	5356530	2257320	341270	456882
2016	5499700	2324193	214062	391116
2017	5594431	2322793	194692	386533
2018	5737236	2336355	206899	351133
2019	5847508	2347433	164533	365828
2020	5917818	2475263	119348	343257
2021	6102195	2390223	38485	312339

Note: classification of full-time equals one when an individual works full-time for one company.
Classification of full-time equals two when an individual works part-time for one company.
Classification of full-time equals three when an individual works full-time for multiple companies in the same industry.
Classification of full-time equals four when an individual works part-time for multiple companies in the same industry.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Table A3 displays the classification of individuals into full-time or part-time workers. Full-time workers can be defined as individuals with one tax certificate in a given year where the 'Total periods in the year of assessment' equals 'Total periods worked in the year of assessment' (indicating that the individual works full time for one company and returns a value of 1 in Table A3). Alternatively, full-time workers are defined as individuals who have multiple tax certificates, all belonging to the same industry, and where the total periods in the year of assessment equal the total periods worked in the year of assessment (indicating that the individual works full time for multiple companies in the same industry and returns a value of 3 in Table A3). Part-time workers are defined as individuals with one tax certificate in the year of assessment where the total periods in the year of assessment do not equal the total periods worked in the year of assessment (indicating that the individual works part time for one company and returns a value of 2 in Table A3). Alternatively, an individual is also classified as a part-time worker when they have multiple tax certificates belonging to the same industry and where the total periods worked in the year of assessment do not sum up to the total periods in the year of assessment (indicating that the individual works part time for multiple companies in the same industry and returns a value of 4 in Table A3).

Our final classification of a full-time worker returns a value of 1 to the full-time variable if they are classified in category 1 or 3 in Table A3. The full-time variable returns 0 otherwise.

For every individual in a given year, we have a full-time variable and imputed industry codes. Next, we sum all labour income across different tax certificates for each tax year.

After this, we keep one observation for each individual, which reports the total labour income earned from all tax certificates, the industry code in which the individual works, and whether the individual works on a full-time or part-time basis.

The date of last access to the data was 29 June 2023.

B Type of certificates by gender

Table B4: Number of IRP5 and IT3(a) certificates issued by gender (2010–21)

	Male				Female			
Year	IRP5	% IRP5	IT3(a)	% IT3(a)	IRP5	% IRP5	IT3(a)	% IT3(a)
2010	4 401 955	29.71	3 961 408	26.74	3 094 313	20.89	3 357 668	22.66
2011	4 576 585	29.48	4 207 027	27.10	3 183 962	20.51	3 559 124	22.92
2012	4 857 337	29.68	4 313 656	26.36	3 511 694	21.46	3 682 756	22.50
2013	4 768 543	28.92	4 496 905	27.27	3 276 363	19.87	3 946 051	23.93
2014	4 938 377	28.94	4 571 560	26.79	3 484 352	20.42	4 071 124	23.86
2015	5 460 747	27.94	5 153 605	26.37	4 259 073	21.79	4 669 955	23.90
2016	5 369 592	28.53	4 921 774	26.15	3 904 567	20.74	4 626 194	24.58
2017	5 618 980	28.35	4 862 442	24.53	4 429 536	22.35	4 907 940	24.76
2018	5 533 952	29.02	4 754 358	24.93	4 092 409	21.46	4 688 578	24.59
2019	5 586 790	29.40	4 591 696	24.17	4 260 269	22.42	4 561 827	24.01
2020	5 538 794	29.50	4 463 262	23.77	4 155 287	22.13	4 619 367	24.60
2021	4 713 567	27.55	4 292 757	25.09	3 607 141	21.08	4 494 988	26.27

Note: IRP5 refers to the number of IRP5 certificates submitted by males or females. % IRP5 shows the per cent of IRP5 certificates submitted by males or females in a given year when the base is all certificates submitted in one year. IT3(a) states the number of IT3(a) certificates submitted by gender and year when the base is all certificates submitted in one year. % IT3(a) shows the per cent of IT3(a) certificates submitted by gender in a year when the base is all certificates submitted in one year. The % shares are calculated for every year. The base for every year is the sum of Male IRP5 %, Male IT3(a) %, Female IRP5 %, and Female IT3(a) % certificates which equal 100.
2008 and 2009 are removed from the table due to the variable indicating the type of certificate being poorly populated.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Table B4 shows the number of tax certificates issued by gender. The table excludes 2008 and 2009 as the variable indicating the type of certificate submitted was poorly populated, yielding low counts. Interestingly, over time IRP5 certificates are the most common certificate type submitted by males; in contrast, females submit more IT3(a) certificates rather than IRP5 forms. Recall, IRP5 records are filed when employees earn more than ZAR2000 and pay tax. From the certificate type that individuals file, women are more likely to make more than ZAR2000 monthly but remain under the tax threshold and work less than 24 hours monthly, whereas men are more likely to have a wage in the tax threshold bounds.

C Additional tables

Table C5: The number of scholarships declared to SARS

Year	Scholarships declared
2008	13 629
2009	13 718
2010	753
2011	781
2012	930
2013	19 364
2014	17 641
2015	17 954
2016	16 827
2017	14 592
2018	20 547
2019	41 327
2020	61 385
2021	22 859

Note: an individual has declared a scholarship or a bursary when the tax certificate has a value corresponding to the following income source codes: 3809, 3859, 3860, or 3815.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Table C6: Labour force participation and unemployment rate by gender

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
							Overall							
LFP % (F)	56.03	54.69	50.58	53.35	53.07	53.72	53.50	55.15	55.56	56.90	56.46	56.92	52.94	54.04
LFP % (M)	69.97	68.08	64.70	65.78	65.94	65.89	65.96	67.69	68.28	68.81	68.24	68.30	64.00	65.25
UE % (F)	21.83	21.59	24.93	23.09	23.35	23.41	24.25	24.84	26.01	25.69	25.88	27.07	25.58	30.27
UE % (M)	17.47	19.56	21.69	19.95	20.43	20.84	21.19	21.18	22.32	22.50	22.77	24.20	23.26	27.46
							Basic education							
LFP % (F)	44.66	43.26	38.98	41.53		41.88	41.67	43.26	43.61	44.46	44.13	44.43	40.93	41.69
LFP % (M)	59.36	57.64	54.11	55.30		54.73	54.92	57.31	58.36	58.26	57.62	57.54	54.17	55.00
UE % (F)	27.70	26.77	30.38	29.08		29.37	31.19	31.06	32.20	31.37	31.09	32.70	29.93	34.48
UE % (M)	21.94	24.28	26.99	24.67		26.25	26.05	25.86	26.68	27.45	27.15	29.04	27.09	31.42
							Advanced education							
LFP % (F)	83.71	83.20	81.51	80.97		80.88	80.91	81.29	80.73	81.21	80.11	79.93	76.73	77.53
LFP % (M)	90.33	89.45	88.29	88.72		89.55	88.39	87.15	86.63	88.24	87.43	86.12	82.50	84.82
UE % (F)	6.59	6.52	7.47	6.65		8.33	9.41	9.22	10.80	13.83	12.68	14.44	14.58	17.17
UE % (M)	4.83	4.92	6.06	6.08		6.20	6.72	7.87	9.87	10.54	10.24	11.62	12.28	14.89

Note: LFP% (F) is the labour force participation rate, female (% of female population ages 15-64). LFP% (M) is the labour force participation rate, male (% of male population ages 15-64). UE% (F) Unemployment, female (% of female labour force). UE% (M) Unemployment, male (% of male labour force).

Source: authors' compilation based on ILO (2022).

Table C7: Proportional of women in managerial positions (2008–21)

Year	% Women in managerial positions
2008	29.6
2009	30
2010	29.6
2011	31
2012	31.1
2013	30
2014	31.1
2015	30.8
2016	31.6
2017	32.1
2018	30.5
2019	30.2
2020	31.4
2021	31.6

Note: the share of women in any managerial position according to the Quarterly Labour Force Survey.

Source: authors' compilation based on StatsSA (2022).

Table C8: Median labour income earned by gender

Income category	Male N	Male income	Male growth	Female N	Female income	Female growth	Cents
			2008				
Poorest 10%	540841	1619.99		382603	1130.04		70
10-25%	811119	10253.99		573809	6897.13		67
25-50%	1351934	28376.94		956391	22337.45		79
50-75%	1351924	63370.51		956406	61208.41		97
75-90%	811159	137741.2		573835	118962.9		86
Richest 10%	540771	312822.3		382558	215436.4		69
Total	5407748	44121.12		3825602	39268.32		89
			2014				
Poorest 10%	638601	2785.074	71.92	503590	1878	66.19	67
10-25%	957890	16039.06	56.42	755280	10246.37	48.56	64
25-50%	1596404	42829.67	50.93	1258841	31649.93	41.69	74
50-75%	1596428	105688.2	66.78	1258855	98077.61	60.24	93
75-90%	957866	240750.2	74.78	755306	216039.5	81.6	90
Richest 10%	638575	527287	68.56	503538	377873.3	75.4	72
Total	6385764	66959.15	51.76	5035410	52881.66	34.67	79
			2021				
Poorest 10%	617545	5303.286	90.42	557923	4042.098	115.23	76
10-25%	926330	24308.88	51.56	832896	15999.53	56.15	66
25-50%	1543852	59792.9	39.61	1390783	44751.19	41.39	75
50-75%	1543860	153157.1	44.91	1390804	133258.5	35.87	87
75-90%	926312	355704.6	47.75	834477	325552.5	50.69	92
Richest 10%	617539	773306.7	46.66	556320	578488.9	53.09	75
Total	6175438	88989.12	32.9	5563203	69480.77	31.39	78

Note: the sample is split by gender, where each gender is organized into six income categories where a corresponding median income is reported. For the year 2008, there is no corresponding growth rate as 2008 is the first year in the data set. Category refers to the percentiles related to the income distribution. N (Male) refers to the number of males in each category. Income (Male) refers to male median income earned by category. Growth (Male) refers to male median income growth from 2008 to 2014 or from 2014 to 2021. N (Female) refers to the number of females in each category. Income (Female) refers to female median income earned by category. Growth (Female) refers to female median income growth from 2008 to 2014 or from 2014 to 2021. Cents is the number of cents that a female earns relative to ZAR1 of male labour income.
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Table C9: Quantile estimates: the relationship between gender and the natural logarithm of real wages

	Q1	Q10	Q20	Q30	Q40	Q50	Q60	Q70	Q80	Q90	Q99
location											
Male	0.248***	0.248***	0.248***	0.248***	0.248***	0.248***	0.248***	0.248***	0.248***	0.248***	0.248***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age	0.134***	0.134***	0.134***	0.134***	0.134***	0.134***	0.134***	0.134***	0.134***	0.134***	0.134***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age^2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Full time	1.577***	1.577***	1.577***	1.577***	1.577***	1.577***	1.577***	1.577***	1.577***	1.577***	1.577***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
scale											
Male	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age	-0.017***	-0.017***	-0.017***	-0.017***	-0.017***	-0.017***	-0.017***	-0.017***	-0.017***	-0.017***	-0.017***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age^2	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Full time	-0.501***	-0.501***	-0.501***	-0.501***	-0.501***	-0.501***	-0.501***	-0.501***	-0.501***	-0.501***	-0.501***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
qtile											
Male	0.334***	0.281***	0.269***	0.260***	0.253***	0.246***	0.239***	0.233***	0.225***	0.215***	0.184***
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Age	0.200***	0.160***	0.150***	0.144***	0.138***	0.133***	0.128***	0.123***	0.117***	0.110***	0.086***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age^2	-0.002***	-0.002***	-0.002***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Full time	3.562***	2.348***	2.066***	1.871***	1.700***	1.541***	1.388***	1.240***	1.068***	0.830***	0.102***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Province included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs	60776758	60776758	60776758	60776758	60776758	60776758	60776758	60776758	60776758	60776758	60776758

Note: standard errors clustered at the individual level are in parentheses. *Male* is coded one for male individuals and zero otherwise. *High skill* refers to the industries that are considered to require a high-skill level to work in as defined by Chen et al. (2022). These industries are: 1) 'Legal Activities', 2) 'Scientific Research and Development', 3) 'Architectural and Engineering activities, Technical testing and Analysis', 4) 'Education Services', 5) 'Human Health Activities', 6) 'Accounting, Bookkeeping, Auditing activities and Tax consultancy', 7) 'Activities of head offices and Management Consultancy Activities', 8) 'Commercial Art and Photography'. *Age* indicates the age of the individual, and Age^2 is the quadratic function of the variable *age*. *Full time* returns one when the individual is a full-time employee and 0 otherwise. *Province* indicates where the individual lives. *Year* controls for the year. *Industry* controls for the industry that an individual works in using a 5-digit SIC code.
* p<0.10, ** p<0.05, *** p<0.01
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Table C10: Quantile estimates: the relationship between gender and the natural logarithm of real wages for high skill sector

	Q1	Q10	Q20	Q30	Q40	Q50	Q60	Q70	Q80	Q90	Q99
location											
Male	0.114***	0.114***	0.114***	0.114***	0.114***	0.114***	0.114***	0.114***	0.114***	0.114***	0.114***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age	0.139***	0.139***	0.139***	0.139***	0.139***	0.139***	0.139***	0.139***	0.139***	0.139***	0.139***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age^2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Full time	1.168***	1.168***	1.168***	1.168***	1.168***	1.168***	1.168***	1.168***	1.168***	1.168***	1.168***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
scale											
Male	0.108***	0.108***	0.108***	0.108***	0.108***	0.108***	0.108***	0.108***	0.108***	0.108***	0.108***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Age	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***	-0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age^2	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Full time	-0.531***	-0.531***	-0.531***	-0.531***	-0.531***	-0.531***	-0.531***	-0.531***	-0.531***	-0.531***	-0.531***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
qtile											
Male	-0.298***	-0.053***	0.007***	0.053***	0.093***	0.128***	0.161***	0.194***	0.228***	0.272***	0.395***
	(0.005)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Age	0.156***	0.146***	0.143***	0.141***	0.140***	0.138***	0.137***	0.135***	0.134***	0.132***	0.127***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Age^2	-0.002***	-0.002***	-0.002***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Full time	3.190***	1.987***	1.691***	1.465***	1.271***	1.097***	0.933***	0.772***	0.606***	0.389***	-0.216***
	(0.007)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Province included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs	3420275	3420275	3420275	3420275	3420275	3420275	3420275	3420275	3420275	3420275	3420275

Note: standard errors clustered at the individual level are in parentheses. *Male* is coded one for male individuals and zero otherwise. *High skill* refers to the industries that are considered to require a high skill level to work in as defined by Chen et al. (2022). These industries are: 1) 'Legal Activities', 2) 'Scientific Research and Development', 3) 'Architectural and Engineering activities, Technical testing and Analysis', 4) 'Education Services', 5) 'Human Health Activities', 6) 'Accounting, Bookkeeping, Auditing activities and Tax consultancy', 7) 'Activities of head offices and Management Consultancy Activities', 8) 'Commercial Art and Photography'. *Age* indicates the age of the individual, and Age^2 is the quadratic function of the variable *age*. *Full time* returns one when the individual is a full-time employee and 0 otherwise. *Province* indicates where the individual lives. *Year* controls for the year. *Industry* controls for the industry that an individual works in using a 5-digit SIC code. * p<0.10, ** p<0.05, *** p<0.01. Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).

Table C11: Linear regression estimates: the impact of gender on real wage earned (excluding IT3(a) tax certificates)

	Dependent variable: Natural log of real income							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male	0.065***	0.079***	0.086***	0.090***	0.114***	0.076***	0.079***	0.115***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
High skill		0.166***	0.225***	0.223***		0.182***	0.179***	
		(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	
Age			0.148***	0.146***	0.091***	0.148***	0.146***	0.091***
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Age^2			-0.002***	-0.002***	-0.001***	-0.002***	-0.002***	-0.001***
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Full time			1.387***	1.385***	1.240***	1.386***	1.384***	1.240***
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Male · High skill						0.110***	0.108***	-0.006**
						(0.00)	(0.00)	(0.00)
Predictive margins: earnings difference by gender and skill level								
Male						0.086***	0.090***	
						(0.00)	(0.00)	
Male · Other skill						0.076***	0.079***	
						(0.00)	(0.00)	
Male · High skill						0.186***	0.187***	
						(0.00)	(0.00)	
Province included	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year included	No	No	No	Yes	Yes	No	Yes	Yes
Industry included	No	No	No	No	Yes	No	No	Yes
Obs	36156363	36156363	36156363	36156363	36156313	36156363	36156363	36156313
Adj. R^2	0.006	0.007	0.298	0.304	0.408	0.298	0.304	0.408

Note: all IT3(a) tax certificates have been excluded from this table. Standard errors clustered at the individual level are in parentheses. *Male* is coded one for male individuals and zero otherwise. *High skill* refers to the industries that are considered to require a high skill level to work in as defined by Chen et al. (2022). These industries are: 1) 'Legal Activities', 2) 'Scientific Research and Development', 3) 'Architectural and Engineering activities, Technical testing and Analysis', 4) 'Education Services', 5) 'Human Health Activities', 6) 'Accounting, Bookkeeping, Auditing activities and Tax consultancy', 7) 'Activities of head offices and Management Consultancy Activities', 8) 'Commercial Art and Photography'. *Age* indicates the age of the individual, and Age^2 is the quadratic function of the variable *age*. *Full time* returns one when the individual is a full-time employee and 0 otherwise. *Province* indicates where the individual lives. *Year* controls for the year. *Industry* controls for the industry that an individual works in using a 5-digit SIC code. * p<0.10, ** p<0.05, *** p<0.01
Source: authors' computation using the SARS tax data (National Treasury and UNU-WIDER 2022).