The wage growth penalty of informal employment

Juan Pablo Siachoque Campos[1]

September, 2022

Abstract

Extensive literature in macro-development has documented a sizable of wage gap between formal and informal workers in developing economies. However, little is know regarding the relation of the wage gap growth with years of tenure of workers in the firm. This documents estimates the tenure wage gap growth between formal and informal workers in Colombia using household survey data between 2007 and 2019. I use two different empirical strategies, (i) Bivariate probit with sample selection and (ii) Multinomial logit with sample selection. The results show that there is a wage gap of 8% after 5 years of tenure and up to 30% after 20 years of tenure.

1 Introduction

An extensive literature in macro-development has documented that low human capital accumulation and poor performance of labor markets in developing economies explain high levels of poverty and inequality. This phenomena is often related to the presence of informal work arrangements, defined as those operating at the margin of the relevant laws and regulations (i.e social security payments), that represent between 20% and 80% of total employment in many developing economies (Perry, 2007).[1] Informal worker are usually characterized by having low education, performing low productive jobs that implies a wage penalty for informal workers compared to their formal peers. Estimates of this penalty, however, vary substantially across countries ranging from 20% to 113% (Ohnsorge and Yu, 2021).

Despite significant magnitude of the wage gap, there is no clear conclusion about its growth and which factors could explain the behaviour. Recent research has found that worker's wage growth over the life cycle in developed economies is steeper than the wage growth of workers in developing economies mainly because of low human capital and learning opportunities in low productive industries (i.e agriculture)(Lagakos et al., 2018). Conversely, there is also evidence that wage growth for workers with high job tenure in developing economies is higher than that described over their life cycle which highlights the importance of labor market match quality (Donovan et al., 2020). The interpretation of these facts is still subject to debate and, in addition, there is no clear connection with informal employment.[2]

Ulyssea (2020) highlights the absence of an analytical framework that conciliates properly human capital and labor market frictions theory with informal employment to explain wage dynamics. A few exception

[1]School of Economics, University of Nottingham, lexjs19@nottingham.ac.uk

[1]This legalistic definition is the dominant approach to define informality for both workers and firms (i.e business registration). An alternative definition is the productivity, which defines informal firms and workers according to low-productivity, firm size or low-quality jobs (for discussion see (Gasparini and Tornarolli, 2009)).

[2]

can be found in Garcia et al. (2015), Bobba et al. (2021) and Bobba et al. (2022) where they study the worker's decision of education and employment either before of after entering the labor market in a life cycle framework. They find that eliminating informal jobs posting through enforcement, increases schooling investments before joining the labor market but decreases the welfare for both workers and firms. After entering the labor market, human capital accumulation and preferences for job amenities in each sector explain the transitions between the informal and the formal sector as well as wage dynamics.

I aim to complement this line of research. In particular, I'm interested in studying the wage growth penalty of informal employment and its relation with job tenure in developing economies. Using Colombia as a study case, I document that there is a significant informal-formal wage tenure gap that growth up to 8% after 5 years of tenure and up to 30% after 20 years of tenure. This finding improves on existing the literature on wage growth and tenure which has overlooked the role of informal employment. Lagakos et al. (2018) does not consider informal wage earners as a factor explaining age-wage growth profiles but include self employment as a robustness check. Moreover, Donovan et al. (2020) use wage earners without a legalistic definition of informality to estimate tenure wage relation.

Using this observation as a starting point, I aim to explore potential interpretation of this significant wage growth differences. The mechanism that explain these facts are usually interpreted by three renown approaches on wage dynamics; namely, human capital, labor market frictions and labor contracts. The first view, builds on the seminal works of human capital accumulation (Becker, 1964), on-the-job training (Ben-Porath, 1967) and learning by doing (Lucas, 2009). This strand states that, features like low human capital accumulation or low productive jobs with limited learning opportunities explain wage growth. The second, is closely related to the seminal work Burdett (1978), Jovanovic (1979) and Jovanovic (1984) where aspects such as low mobility in the labor markets influence the misallocation of workers on skilled jobs and limits wage growth. The third, refers to Lazear (1979) which highlights the importance compensation schemes in certain work arrangement like long versus short term influence wage growth.

For the specific case of wage growth and tenure in developing economies, Donovan et al. (2020) elaborates from a learning and job ladder framework. From the learning models perspective, workers and firms are imperfectly informed about match productivity but learn more by producing so if they learn that a match is unproductive, they endogenously choose to separate and conversely remain in the match if productive (Jovanovic, 1979). From the job ladder models perspective, workers receive attractive wage offers outside their current job that induces workers to quit (Burdett and Mortensen, 1998). Both the empirical evidence on the steepness of tenure-wage profiles as well as the theory to explain it, lacks a clear classification between formal and informal workers to informal-formal wage gap.

The document is organized as follows. Section 2, provides a brief review of literature on informality. Section 3 describes the data and empirical strategy. In Section 4, I discuss a potential driver of this wage gap and in Section 5 discuss how this fact could be a potential path for future research.

2 Related literature

This research relates to different strands of literature that study informality from both the perspective of firms and workers. The majority of research on informality has been focused on the firms, where there three renown approaches: the survival (or dual), the exclusion view or the parasite (or rational) view (Ulyssea, 2018). The first, refers to informality as a low productivity phenomena where firms are too unproductive to ever become formal even in the absence of entry cost. The second view, corresponds to potentially productive firms that are excluded from formality because of high entry costs. The third, states that once entry barriers are removed (or are low) productive firms weight cost and benefits to become formal and select to remain informal.

Recent evidence shows that these approaches can complement each other, particularly because formal and informal firms coexist within defined industries producing similar products and where their productivity overlap substantially (see Ulyssea (2018), Meghir et al. (2015)). In addition, literature has emphasized that the share of informal firms decline as firms grow larger, potentially indicating that factors such as enforcement or credit access increase the opportunity cost to remain informal. In this regard, a less explored view is that the share of informal workers within formal firms declines with firm size (Cisneros-Acevedo, 2022).[3] This last observation is key to understand labor market outcomes for formal and informal workers while working on a formal firm.

Among those labor market outcomes, the literature has documented a sizable informal-formal wage gap even controlling by observable characteristics and increases when different educational levels and for different quantiles of the wage distribution (Quintin (2008), Bargain and Kwenda (2014)). However, Ulyssea (2018) finds no wage gap evidence using matched employer-employee data for Brazil on both formal and informal after controlling for firm fixed effects. A similar finding is shown for Ecuador, where they claim that only larger firms, which are more likely to be formal, pay higher wages and small firms that are usually informal pay low wages thus the informal formal wage gap is a reflection of firms size gap (El Badaoui et al., 2010).

Ulyssea (2020) suggest that this finding informal-formal wage gap and firm size wage gap is key because if there is an underlying matching mechanism between firms and workers that, after controlling worker and firm characteristics, point to a strong self-selection effect on wage gap between workers with same observable characteristics that differ only on firm size or that conditional on skills, formal and informal workers could perform the same tasks within the firm that explain why in some cases there is no wage gap.

Regarding the tenure related literature, wage growth is usually interpreted as job specific accumulation of human capital Becker (1964), compensation schemes to improve efficiency of workers and increase wages (Lazear, 1979) or deferred compensation schemes to provide incentives to workers through higher wages Salop and Salop (1976). In this, line of research and more recently Donovan et al. (2020) argues that it is unlikely that wage growth in developing countries could be explained by job-specific accumulation because the prevailing low levels of human capital but it could be explained because of the selection of

[3] Other strand of literature documents that firms in developing economies grow less as they age and that stagnant firms tend to survive longer compared to developed economies (see Akcigit et al. (2021), Eslava et al. (2019)).

workers and firms into quality matches that survive to longer tenure. This perspective abstracts from a explicit definition of informal employment for wage earners and how it impacts tenure-wage profiles.

However, other types of work arrangements and human capital accumulation connection have been studied in the literature to explain wage growth such as short-term flexible labor contracts and gig economy related employment. The main message of this strand of literature, is that there are differences of general, specific and transferable skills gains depending on the type of work arrangement. These work arrangements offer and alternative to unemployment and human capital depreciation Guvenen et al. (2017) but it could be negative if it induces instability of employment García-Pérez et al. (2019) and lower on the job training opportunities Bratti et al. (2021). Overall, the evidence shows that there is a wage penalty in short term contracts compared to other long term work arrangements.

3 Empirical strategy

3.1 Data

I use micro-data from the Colombian Household Survey (GEIH) collected by the National Statistics Office (DANE) from 2007 to 2019. The survey consist of cross sections with national coverage (urban and rural) corresponding to the largest five regions, 23 states, 13 main cities with their metropolitan areas and 11 intermediate cities. The repeated cross-section approximately covers 230,000 individuals per year and compiles information about size, structure and characteristics of the labor force, employment, unemployment, inactivity status and quality of employment. The data measures individual's monetary income and in-kind, housing, access to public services and programs and social security.

The data permits to classify workers by type such as wage earner in private or public sector, non wage earner or self employed. Due to the wage focus, I use only wage earners in the private sector and classify them as formal or informal using the legalistic approach standard in the literature (Perry, 2007).[4] In particular, if a wage earner declares not to have social security benefits or having the benefits but not payed by the employer, they are classified as an informal wage earner worker (see Annex 6.1). Table 1, depicts that wage earners represent 43% of total employment and the share of informal wage earner workers is 30.2%.

Informal worker's demographic exhibit particular patterns according to age, educational level and other characteristics. The figure in the Annex 6.2 depicts that informality has a U-shaped behaviour according to age because it is higher among young and old workers (40%) for ages between 20 and 25 years and between 60 and 65 years, while the level of informality for other age groups is 22% on average. Additionally, informality decreases as the years of education increase where the share of informal workers with no education is nearly 80%. Moreover, informality is higher in small cities (41% vs 31%) and is higher for men than for women (34% vs 31%). Finally, by small size firms measured by number of employees informality is 80% and higher for workers with low tenure (25%).

[4]Using wage earners only is convenient as is a direct payment in a labor contract between a firm and a worker that is broadly reported in the survey. In contrast, the income of the self-employed is reported as gains which includes both labor and capital income and makes it difficult to asses accurately wage dynamics. In addition, although public workers receive a wage the underlying dynamics is based on predefined compensation schemes.

Table 1: Distribution of employment by type and status, 2007-19, all ages

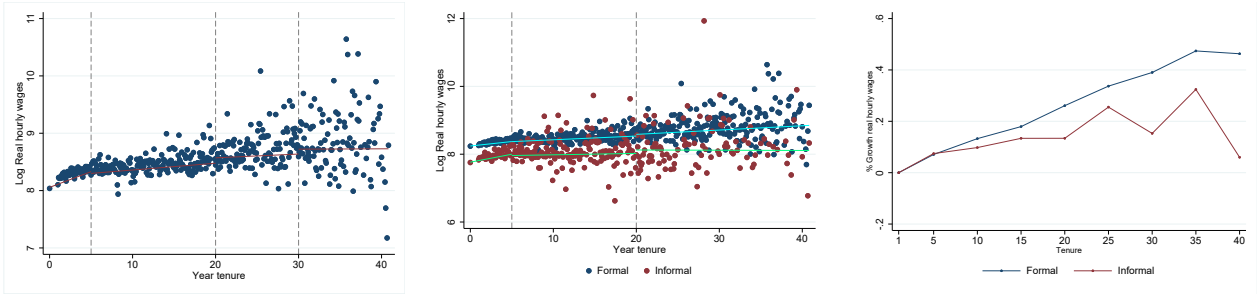
A. Employment		
	Observations	Share of employment
Wage earners	1,076,594	43.81
Self employed	990,640	40.31

B. Formal/informal status		
	Share of employment	Share of wage earner
- Formal	30.6	69.83
- Informal	13.22	30.16

Source: Statistics office and author's elaboration. Computations with survey weights.

On this last observations, the variable tenure in the survey refers to, how long in months has the individual worked in the current establishment continuously. Moreover, there is a positive relationship between tenure and wage level and growth. In Figure 1(a)(b), I depict a binscatter plot for the logarithm of real hourly wages and years of tenure for overall wage workers and by employment status.[5] In both cases of status, there is a positive relationship of the level of wages and tenure that increases it's dispersion for years beyond 20. However, the dispersion increases earlier for informal workers more less by year 10 and amplifies afterwards. The growth of wages is depicted in Figure 1(c) which describes no significant gap of wage growth for the first 5 years but grows to nearly 10% by 20 years of tenure.

Figure 1: Level and growth of hourly real wages and tenure



(a) Bin-scatter - All workers (b) Bin-scatter Formal-Informal workers (c) Wage growth and tenure

Source: Statistics office and author's elaboration. Variables are computed with survey weights.

[5]For the monthly wage measure reported in the survey, I remove other labor income such as bonuses, tips and commission if any. To deflate the nominal wage, I use the annual rate of inflation for the reference month. In addition, I multiply hours worked by four as they are originally reported as hours per week. Finally, wages refer to the primary job, abstracting for the a second job because represents less than 4% of total employment.

3.2 Estimations

Most estimations of informal-formal wage gap use the Mincer wage equation approach and including a dummy variable that defines informality and controlling for additional worker and firm characteristics. Using this framework as a starting point, I discuss the issue of selection in this approach as well as two empirical strategies to reduce selection bias. In addition, to explore the role of tenure on wage growth, I use the Topel (1991) approach to study wage growth.

3.2.1 Wage profiles[6]

To explain the empirical strategy, first consider the following regression for overall wage earners worker i:

$$log(w_i) = \beta \mathbf{X}_i + \varepsilon_i, \tag{1}$$

Where w_i is the wage and $\mathbf{X}_i$ includes level of schooling and other firm and worker characteristics. Under the assumption of exogeneity $E[\varepsilon_i|\mathbf{X}_i] = 0$, so estimating Equation 1 is consistent. However, there is an issue with selection bias, in particular, there could be two potential sources of selection: (i) individual i selects to participate in the in the labor market as wage earner and (ii) the individual i selects to be either formal or informal. The way we think about the sources of selection is key, because it shapes the decision of individuals of joining either sector and is directly connected with the three views of informality: survival, parasite and exclusion.

Moreover, wages are only observed for individual i when they select to work as wage earners but we cannot observe the wage of individual i for all other possible choices such as inactive, unemployed or self employed. Now consider the decision to participate in the labor market as a wage earner as a treatment. Let S_i denote a treatment that takes the value of 1 if the worker is wage earner and 0 otherwise. Thus, we have two potential outcomes with observed and unobserved determinants for individual i:

$$log(w_{0,i}) = \alpha_0 \mathbf{X}_i + u_{0,i}, \quad if \quad S_i = 0$$

$$log(w_{1,i}) = \alpha_1 \mathbf{X}_i + u_{1,i}, \quad if \quad S_i = 1$$

Where the first outcome is the wage when the individual is not treated with $E[u_{0,i}|\mathbf{X}_i] = 0$ and the second outcome is the wage when the individual is treated with $E[u_{1,i}|\mathbf{X}_i] = 0$. Thus, the observed wage is $log(w_i) = Slog(w_{1,i}) + (1 - S)log(w_{0,i})$. In a cross section it is not possible to know both $log(w_{0,i})$ and $log(w_{1,i})$ for any i. For those going to work as wage earners, we cannot observe $log(w_{0,i})$; for those not working as wage earners, we cannot observe $log(w_{1,i})$. The individual treatment effect is thus the difference between the two outcomes:

$$\Delta_i = log(w_{1,i}) - log(w_{0,i}) = \alpha_1 \mathbf{X}_i - \alpha_0 \mathbf{X}_i + \underbrace{u_{1,i} - u_{0,i}}_{\text{Unobserved}}$$

Note that α_0 and α_1 could or not be common between treated or not treated individuals this would point out to heterogeneous characteristics of individuals. In addition, heterogeneity could also come from

[6] An alternative view to study wage growth could be as a age-wage profile, in Annex 6.5, I do a simple procedure to estimate wage growth over the life cycle.

unobserved or idiosyncratic factors $u_{j,i}$.[7] This idiosyncratic unobserved factors mean that even if we observe, for instance, individual i with high education and expected be in the labor market it could be the case that there is some unobserved idiosyncratic factor $u_{0,i}$ that is large enough so that individual i is not treated.

Thus, a treatment can be understood as a choice or selection of each worker that takes place because an underlying decision rule that makes individual i take or not the treatment. A way to interpret this is the difference between offered wages and reservations wages, if a wage offer does not meet or exceed the worker's reservation wage, then the worker's utility is maximized choosing other occupation. Let this underlying decision be expressed in term of individual utility:

$$U_i = \gamma \mathbf{Z}_i + \nu_i$$

Where $\mathbf{Z}_i$ is a set of observable instruments that explain the incentives to becoming a wage earner and $E[\nu_i] = 0$. If utility $U_i > 0$ the worker selects to be a wage earner. We assume that $\mathbf{X}_i$ is a subset of $\mathbf{Z}_i$. In practice we don't observe the utility but a dummy variable S_i. Keeping this in mind, the observed wage can now be rewritten as:

$$log(w_i) = Slog(w_{1,i}) + (1 - S)log(w_{0,i}) \tag{2}$$

$$log(w_i) = S\Big[log(w_{1,i}) - log(w_{0,i})\Big] + log(w_{0,i}) \tag{3}$$

$$log(w_i) = S\Big[\big(\alpha_1 - \alpha_0\big)\mathbf{X}_i + u_{1,i} - u_{0,i}\Big] + \alpha_0 \mathbf{X}_i + u_{0,i} \tag{4}$$

$$log(w_i) = \alpha_0 \mathbf{X}_i + S\underbrace{\big(\alpha_1 - \alpha_0\big)\mathbf{X}_i}_{\Delta_i} + \underbrace{S\big(u_{1,i} - u_{0,i}\big) + u_{0,i}}_{\varepsilon_i} \tag{5}$$

We can't really know Δ_i (the individual treatment effect) because that would mean that we can see both outcomes for i, instead we see an average effect $E[\Delta_i]$. However there is a problem because the treatment of joining the labor market as wage earner also determines the error component ε and thus $E[\varepsilon S_i] \neq 0$. If the selection into treatment is not random but is determined by a decision rule, the estimation of the average treatment $E[\Delta_i]$ is biased, in particular:

$$E[log(w_i)|S_i = 1] - E[log(w_i)|S_i = 0] = \underbrace{E[\Delta_i]}_{\text{ATE}} + \underbrace{E[u_{1,i}|S_i = 1] - E[u_{0,i}|S_i = 0]}_{\text{Selection Bias}}$$

In this case with selection bias there is no exogeneity and a ordinary least squares estimate will be inconsistent. Moreover, if we include to the informal workers to the potential analysis framework there is an additional source of selection bias. In particular, the wages cannot be observed for both formal and informal for each individual i. That is, as in the previous case, if individual i selects to be a formal worker it cannot be observed an informal wage and vice versa.

[7] Heckman and Li (2004) elaborates for heterogeneous effects effects across i meaning $\gamma_{j,i}$.

To consider this sources of selection, I use two separate strategies (i) Bivariate Probit with selection bias which considers a two stage approach where the first refers to wage earner selection, and the second stage refers to the informal and formal selection and (ii) the Multinomial Logit with selection bias, considers selection of all occupational choices in the first stage either wager earner informal or formal or self employed. Afterwards, I use these procedures to estimate the wage regression in the spirit of (Heckman, 1979) who demonstrates that the selection bias can be seen as an omitted variable bias in a standard OLS regression that once included the estimated parameters are consistent.

Bivariate Probit with Selection: The underlying idea of this approach is that informal or formal status is only observed if the worker selects to be a wage earner. In this case, the outcome is whether the individual is informal or not and the selection equation is the whether the individual chooses to be wage earner or not.[8]

$$status_i = \gamma_1 \mathbf{Z}_{1,i} + \varepsilon_{1,i}, \qquad status_i = formal_i, informal_i \tag{6}$$

$$wage.earner_i = \gamma_2 \mathbf{Z}_{2,i} + \varepsilon_{2,i}, \tag{7}$$

Equation 6 is the outcome equation where the dependant variable $status_i$ could be either a $formal_i$ or $informal_i$ dummy variable that takes the value of 1 if the worker is classified in that particular status and zero otherwise. $\mathbf{Z}_{1,i}$ is the set of instruments influencing individual's decision of becoming informal or formal such as demographics such as age (and a squared term), years of education, a dummy for gender, a dummy variable for city size, marital status, time of transition between the previous and the current job, tenure in current establishment express as a third degree polynomial, firm size and industry dummies.

The Equation 7 is the selection equation where $wage.earner_i$ is a dummy variable that takes the value of one if the worker selects to be wage earner and zero otherwise (zero includes self employed, inactive or unemployment). $\mathbf{Z}_{2,i}$ are the set of instruments related to the decision rule to participate in the labor market. This instruments include, the number of dependants within the household; that is, those individuals under 18 years old and above 65 years old, and a dummy indicating whether the house is fully or partially owned or not. Moreover, notice that there is overlap between the variables in both equations, as some characteristics influence both occupational choice and employment status. This means that $\mathbf{Z_{1,i}}$ is included in the selection equation but the set $\mathbf{Z_{2,i}}$ is excluded in the outcome equation[9].

As stated previously, there is a latent process related to the utility of the individual i such that induces the selection of wage earner as occupation. Whenever this underlying process yields positive utility we can observe a wage earner worker in the sample. The decision rule is the following:

$$wage.earner_i = \begin{cases} 1, & \text{if} \quad U_i = \gamma_2 \mathbf{Z}_{2,i} + \varepsilon_{2,i} > 0 \\ 0, & \text{if} \quad U_i = \gamma_2 \mathbf{Z}_{2,i} + \varepsilon_{2,i} \leq 0 \end{cases}$$

[8] A similar approach is used in García and Badillo (2018), however they use a framework of bivariate probit with partial observability.

[9] This is refereed as the exclusion restriction where due to the number of dependants determines the participation in the labor market but does not determine the informal or formal status.

In addition, assume that the disturbances are distributed $\varepsilon_{1,i} \sim N(0,1)$, $\varepsilon_{2,i} \sim N(0,1)$ and the correlation between them is $cov(\varepsilon_{1,i}, \varepsilon_{2,i}) = \rho$. The existence of a correlation $\rho \neq 0$ suggest that employment status is conditioned to the probability of selecting or not $wage.earner_i$ as an occupational choice. In this setting, the probability of being in any status is given by a normal joint density probability ϕ_2:

$$\phi_2(\varepsilon_1, \varepsilon_2) = \frac{1}{2\Pi\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}} \exp\left[-\frac{1}{2}\frac{(\varepsilon_1^2 + \varepsilon_2^2 - 2\rho\varepsilon_1\varepsilon_2)}{(1-\rho^2)}\right]$$

And a cumulative density function:

$$\Phi_2(\varepsilon_1, \varepsilon_2) = \int_{\varepsilon_1}\int_{\varepsilon_2} \phi_2(\varepsilon_1, \varepsilon_2, \rho) d\varepsilon_1 d\varepsilon_2$$

If $\rho = 0$, then the two errors are independent and the Φ_2 reduces to two separate standard normal distributions. If $\rho \neq 0$, then the two errors are correlated and the probability of formal or informal status will depend on the probability of selecting occupation wage earner. If $\rho = 1$, then the two variables are essentially the same. The probabilities involved in this bivariate setting can be computed in the following way:

When an individual i selects wage earner as occupation this probability can be expressed as:

$$Pr(\varepsilon_{2,i} > -\gamma_2 \mathbf{Z}_{2,i}) = 1 - Pr(\varepsilon_{2,i} \leq -\gamma_2 \mathbf{Z}_{2,i}) = 1 - \Phi(-\gamma_2 \mathbf{Z}_{2,i})$$

And equivalently the probability of not selecting wage earner as occupation is $\Phi(-\gamma_2 \mathbf{Z}_{2,i})$. The probability of being, say informal, conditional on selecting wage earner:

$$Pr(wage.earner_i = 1, informal_i = 1) \quad = \quad \int_{-\infty}^{\varepsilon_1}\int_{-\infty}^{\varepsilon_2} \phi_2(\varepsilon_1, \varepsilon_2, \rho) d\varepsilon_1 d\varepsilon_2 \quad = \quad \Phi_2(\gamma_1 \mathbf{Z}_{1,i}, \gamma_1 \mathbf{Z}_{2,i}, \rho)$$

Moreover, the probability of not being informal conditional on selecting wage earner:

$$Pr(wage.earner_i = 1, informal_i = 0) \quad = \quad \Phi(\gamma_1 \mathbf{Z_1}_i) - \Phi_2(\gamma_2 \mathbf{Z}_{2,i}, \gamma_1 \mathbf{Z}_{1,i}, \rho)$$

The same framework applies when the status is formal. Moreover, the estimation procedure can be achieved using maximum likelihood method in two steps: the first stage estimates a probit model for Equation 7, and the second stage estimates a probit model considering the conditional probabilities. I estimate a separate Bivariate model for each status (formal and informal) and compare them with a simple univariate model.[10]

[10] As in the standard probit model, each observation contributes some combination of $Pr(informal_j = 1)$ for $j = 1, 2$ to the likelihood function depending on their specific values. Thus, the log-likelihood is just a sum across all possible transition probabilities, which in this case they are three because recall that if we have $wage.earner_i = 0$ we cannot see $status_i$ neither formal or informal. Therefore the possible combinations when status is informal are $Pr(wage.earner = 0)$,$Pr(wage.earner = 1, informal_i = 0)$,$Pr(wage.earner = 1, informal_i = 1)$. Thus the log-likelihood function is given by:

$$\begin{aligned} ln\mathbf{L} = \sum_{i=1}^{N} [&(informal_i)(wage.earner_i) ln(\Phi_2(\gamma_1 \mathbf{Z}_{1,i}, \gamma_1 \mathbf{Z}_{2,i}, \rho)) \\ &+((wage.earner_i)(1 - informal_i) ln(\Phi(\gamma_2 \mathbf{Z_2}_i) - \Phi_2(\gamma_2 \mathbf{Z}_{2,i}, \gamma_1 \mathbf{Z}_{1,i}, \rho))] \\ &+(1 - wage.earner_i) ln(\Phi(-\gamma_1 \mathbf{Z}_{1,i})) \end{aligned}$$

Table 2: Estimates of bivariate probit model for each employment status

	Univariate Probit		Bivariate Probit Informal		Bivariate Probit Formal	
			First Stage	Second Stage	First Stage	Second Stage
Dep. Var	Informal	Formal	Wage Earner	Informal	Wage Earner	Formal
	(1)	(2)	(3)	(4)	(5)	(6)
Tenure	-0.14***	0.14***	-0.05***	-0.12***	-0.04***	0.13***
Tenure^2	0.01***	-0.01***	0.00***	0.01***	0.00***	-0.01***
Tenure^3	-0.00***	0.00***	-0.00***	-0.00***	-0.00***	0.00***
Age	-0.03***	0.03***	0.02***	-0.02***	0.01***	0.01
Age^2	0.00***	-0.00***	-0.00***	0.00***	-0.00***	0.00
Education years	-0.06***	0.06***	-0.02***	-0.05***	-0.01***	0.05***
Gender (1=Male)	-0.14***	0.14***	0.07***	-0.15***	0.05***	0.12***
Firm size (No. employees)						
2 - 5	0.12***	-0.12***	2.83***	-0.93***	2.75***	-2.17***
6 - 20	-0.76***	0.76***	3.89***	-1.90***	3.81***	-2.08***
20 - 50	-1.47***	1.47***	4.15***	-2.53***	4.08***	-1.58***
50 >	-2.09***	2.09***	4.14***	-3.08***	4.12***	-1.02***
Industry						
Agriculture and Mining	-0.27***	0.27***	0.51***	-0.26***	0.45***	0.00
Manufacturing	-0.02**	0.02**	0.41***	0.05**	0.31***	-0.15***
Construction and retail	-0.11***	0.11***	0.32***	-0.19***	0.31***	0.05*
Transportation and trade	-0.29***	0.29***	0.19***	-0.32***	0.20***	0.22***
Private Services	-0.26***	0.26***	0.18***	-0.23***	0.12***	0.14***
City	-0.14***	0.14***		-0.22***		0.18***
Partner	-0.07***	0.07***		-0.10***		0.08***
Time in job transition	0.01***	-0.01***		0.00***		-0.00
Instruments						
1 - 3			0.01		-0.06**	
4 - 5			-3.62***		-2.45***	
Home ownership			-0.10***		-0.02*	
Rho			-0.48***		-1.44***	
Constant	2.53***	3.50***	-2.89***	3.50***	-2.84***	1.07***
Observations	635,194	252,631	252,631	252,631	190,413	156,724
AIC	431284.4	442121.6		200716.7		77.6781
BIC	431511.6	442326		201144.8		474.4743
Wald test						
exlusion var χ^2				407.84		264.29
p-val				0.00		0.00

Source: Statistics office and author's elaboration. *** p<0.01, ** p<0.05, * p<0.1. The table depicts the coefficients of the estimation.

Table 2 show the results of the estimation. First note that, the first stage (Columns (3),(5)) the instruments are significant in most cases. There is a higher likelihood of joining the labor market as a wage earner when the number of dependants is between one to three but drops as the number of dependants increase. This is probably explained by the trade off between working in a full time job and being able to take care children or the elder of the household as mentioned by García and Badillo (2018). In addition, owning a house fully or partially has a negative effect on choosing to be wage earner, potentially suggesting a wealth effect that provides an incentive not to chose this occupation. However, there is a problem with using income or wealth variables at this stage due to the simultaneous bias that could arise between work and income. Additionally, I test that all exclusion variables are statistically equal to zero. The Wald test suggest rejection at one percent that all variables have jointly no effect on the employment equation supporting the exclusion restriction is relevant to determine that the individual selects wage earner occupation.

Regarding the second stage, there are important differences between the univariate and bivariate estimations for both work status signaling an important selection effect (Columns (1)(4) and (2)(6)). In the case of informal status, there is a significant difference in tenure; there is a lower likelihood of being informal

when tenure increases compared to the univariate case. Indeed this is an observation on the data where there is fewer informal workers for higher years of tenure. Other result suggest that as age increase there is a lower likelihood of being informal to that of being formal (this relates to the informality among young workers). Moreover, lower education reduces the likelihood of selecting informal work and the opposite happens for formal workers. Finally, women workers, living in small cities without a partner are more likely to be informal workers.

In addition, there is a significant difference in the firms size likelihood of being informal suggesting a upward bias, this is closely connected tho the fact that informal work is usual in small size firms of less that 20 employees. This factor has the most significant change for informal and formals estimation between univariate and bivariate framework. Finally, by industry there is a significant higher likelihood of selecting formality in most industries except in manufacturing.

After estimating the bivariate model with selection, we can compute the predicted probabilities of Equation 6 for both formal and informal status and follow the Heckman (1979) approach to estimate through OLS the wage profiles for formal and informal workers controlling with the inverse Mills ratio. Recall that our wage equation is

$$log(w_i^j) = \beta \mathbf{X}_i^j + \epsilon_i^j, \qquad j = formal_i, informal_i \tag{8}$$

Consider that observing a wage is conditional to observing a formal or informal worker, that is:

$$\begin{aligned} E[log(w_i^j) \mid j = 1] &= E[\beta \mathbf{X}_i + \epsilon_i \mid \hat{\gamma}_1 \mathbf{Z}_{1,i} + \hat{\varepsilon}_{1,i}], \quad j = formal_i, informal_i \\ &= \beta \mathbf{X}_i + E[\epsilon_i \mid \hat{\gamma}_1 \mathbf{Z}_{1,i} + \hat{\varepsilon}_{1,i}] \end{aligned}$$

We assumed that the distribution of error terms in the bivariate framework are normal, thus.

$$= \beta \mathbf{X}_i + \underbrace{\rho \sigma_\epsilon \frac{\phi\left(\frac{\hat{\gamma}_1 \mathbf{Z}_{1,i}}{\sigma_{\varepsilon_1}}\right)}{\Phi\left(\frac{\hat{\gamma}_1 \mathbf{Z}_{1,i}}{\sigma_{\varepsilon_1}}\right)}}_{\text{Inverse Mills ratio}}$$

In practice, I estimate the following regression similar to the specification of Topel (1991) to retrieve a wage tenure profile for each type of worker:

$$log(w_i^j) = \alpha + \sum_{k=1}^{3} \varphi_k tenure_i^k + \beta X_i + \theta MillR_i^j + \varepsilon_i^j, \quad j = formal, informal. \tag{9}$$

Table 3 depicts the point estimates of the OLS estimation of Equation 9. First, note that the Inverse Mills ratio in both cases is positive and statistically significant in both formal and informal regression (2) and (4) but the magnitude is higher for formal wage regression. Furthermore, tenure coefficients have an a upward bias for informal and downward bias for formals, meaning that there is a premium effect of tenure while working for longer formally. A similar effect is found on age and education where it is higher for formals and lower for informal workers.

The most salient change of coefficient estimation is that of firm size in both informal and formal regressions. In the case of informal workers, the selection controlled version of OLS describes a significant upward bias; the estimates for firm size change from positive to negative. This results, relates somewhat to the findings of El Badaoui et al. (2010) where there is a wage closely connected to a firm size gap.

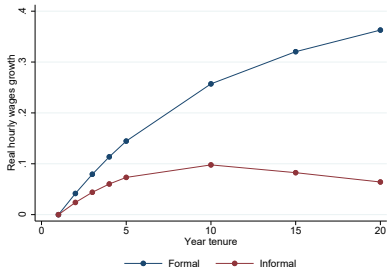
Table 3: Ordinary least squares estimates

	Informal		Formal	
Dep. Var - log(wage)	(1)	(2)	(3)	(4)
	-	Sel. correction		Sel. correction
Tenure	0.05***	0.03***	0.02***	0.05***
$Tenure^2$	-0.00***	-0.00***	-0.00***	-0.00***
$Tenure^3$	0.00***	0.00***	0.00***	0.00***
Age	0.02***	0.01***	0.01***	0.01***
Age^2	-0.00***	-0.00***	-0.00	-0.00
Education years	0.04***	0.03***	0.07***	0.08***
Gender	0.12***	0.09***	0.06***	0.08***
Firm size (No. employees)				
2 - 5	0.15***	-0.04*	0.07*	-0.43***
6 - 20	0.31***	-0.05*	0.16***	-0.31***
20 - 50	0.34***	-0.16***	0.20***	-0.11***
50 >	0.41***	-0.25***	0.29***	0.09***
Industry				
Agriculture and Mining	0.34***	-0.16***	0.20***	-0.11***
Manufacturing	0.41***	-0.25***	0.29***	0.09***
Construction and retail	0.02	-0.03	-0.13***	-0.15***
Transportation and trade	0.08***	0.06***	0.08***	0.01*
Private Services	-0.10***	-0.14***	-0.15***	-0.15***
Inv. Mills Ratio		0.35***		0.76***
Constant	6.69***	7.12***	6.72***	6.59***
Observations	190,413	156,724	512,783	439,653

Source: Statistics office and author's elaboration. *** p<0.01, ** p<0.05, * p<0.1.

Finally, to retrieve the wage profile, I compute the predicted wage growth for tenure years 1 to 5, 10 and beyond and compute the the growth with respect to the first year of tenure. Figure 2 shows that by five years of tenure there is a wage growth gap of 8% and then grows to 16% by year 10 and to 30% by year 20.

Figure 2: Growth of hourly real wages and tenure with Heckman-probit procedure



Source: Statistics office and author's elaboration.

Multinomial Logit with Selection Bias: When selection is over a large number of exclusive choices (i.e. informal and wage earner or self employment), the multinomial logit specification is convenient. Bourguignon et al. (2007) documents a survey of different approaches in a unified framework. For simplicity, I use the framework formal and informal choice and self employment in the multinomial framework. In this approach the equation of interest and a selection equation:

$$log(w_{1,i}) = \beta_1 \mathbf{X}_i + \varepsilon_{1,i},$$
$$ocup_k = \gamma_k \mathbf{Z}_k + \eta_k, \qquad k = 1, ..., M. \tag{10}$$

Where the subscript 1 is the first occupations choice assumed to be formal wage earner. Assume that $E[\varepsilon_1|\mathbf{X}_i, \mathbf{Z}_i] = 0$ and $V[\varepsilon_{1,i}|\mathbf{X}_i, \mathbf{Z}_i] = \sigma^2$, k is the occupation choice of the individual i among M categories. Thus, the individual i chooses the alternative 1 when it yields the maximum utility among all choices.

$$\begin{cases} ocup_1 & if \quad U_1 > \max(U_k) \quad \forall \quad k \neq 1 \\ ocup_k & if \quad U_1 < \max(U_k) \quad \forall \quad k \neq 1 \end{cases}$$

Now define an error term u_1

$$u_1 = \max_{k \neq 1} (U_k - U_1)$$
$$u_1 = \max_{k \neq 1} (ocup_k - ocup_1)$$
$$= \max_{k \neq 1} (z\gamma_k + \eta_k - z\gamma_1 - \eta_1)$$

Note that if occupation 1 is chosen, it means that the choice yields the maximum utility among the alternatives and therefore the difference $U_k - U_1$ is negative. In addition, assume that η_k is iid with Gumbel density function $g(\eta) = \exp(-\eta - e^{-\eta})$ and cumulative $G(\eta) = \exp(-e^{-\eta})$. This assumption leads to the multinomial logit model with a probability of occupation formal wage earner given by

$$P(u_1 < 0 \mid z) = \frac{\exp(z\gamma_1)}{\sum_k \exp(z\gamma_k)}$$

Consistent maximum likelihood estimates of each (γ_k) parameter can be computed. The problem is to estimate the parameter vector β_1 when there is correlation between outcome disturbances and occupational choice disturbances are correlated $cov(\varepsilon_1, \eta_k) \neq 0$. In this case, OLS estimates of β_1 would not be consistent. Thus, if we consider the conditional density of the outcome error component ε_1 on u_1

$$E(\varepsilon_1 \mid u_1 < 0, \Gamma) = \int\int_{-\infty}^{0} \frac{\varepsilon_1 f(\varepsilon_1, u_1 \mid \Gamma)}{P(u_1 < 0 \mid \Gamma)} du_1 d\varepsilon_1 = \lambda(\Gamma) \qquad \Gamma = \{z\gamma_1, z\gamma_2, \ldots, z\gamma_M\}$$

where $f(u_1, \varepsilon_1 \mid \Gamma)$ is the conditional joint density of u_1 and ε_1 and the function Γ is simply the instruments on each occupational choice. Thus, the term λ is the selection bias correction term that is a function of instruments $\mathbf{Z}$. Now considering the probability P_j as the probability that any alternative j is preferred among all k alternatives:

$$P_j = \frac{\exp(z\gamma_j)}{\sum_k \exp(z\gamma_k)}$$

Bourguignon et al. (2007) states that there is a unique function μ that can substitute $\lambda(\Gamma)$ due to the

correspondence of the M components of Γ and the M corresponding probabilities.

$$E\left(\varepsilon_1 \mid u_1 < 0, \Gamma\right) = \mu\left(P_1, \ldots, P_M\right)$$

Therefore, consistent estimation of β_1 can be achieved regressing the following using OLS:

$$log(w_{1,i}) = \beta_1 \mathbf{X}_1 + \mu\left(P_1, \ldots, P_M\right) + \nu_1 \tag{11}$$

where $E[\nu_1|\mathbf{X}_1] = 0$. The framework has a problem with dimensionality whenever the alternative set M is large. Bourguignon et al. (2007) claims that there are two main ways to deal with dimensionality: (i) restrictions over the covariance matrix of the error terms (ε, η) and (ii) linearity assumptions. For the moment, I use the Dahl (2002) approach which restricts the set of probabilities in $\mu\left(P_1, \ldots, P_M\right)$ to a chosen subset S of particular interest, making the hypothesis that this subset contains all the relevant information. This is called the Dahl's index sufficiency assumption, thus:

$$f\left(u_1, \varepsilon_1 \mid \Gamma\right) = f\left(u_1, \varepsilon_1 \mid P_{j,j=1,\ldots,M-1}\right) = f\left(u_1, \varepsilon_1 \mid P_{j,j\in\Theta}\right), \quad \Theta \quad \subset \{1, \ldots, M-1\}$$

So the estimated equation under Dahls approach becomes:

$$log(w_{1,i}) = \beta_1 \mathbf{X}_i + \mu\left(P_{j,j\in\Theta}\right) + \nu_1$$

There are two main challenges mentioned by (Dahl, 2002): (i) testing whether the index sufficiency assumption is not generally feasible and (ii) the reduction of dimensionality is made at the cost of a restrictive assumption on the correlation structure of the error terms. A special case proposed by Dahl is to make the hypothesis that the probability of alternative 1 is the only information needed which in this case is fits the our aim to find the probability of selecting formal (or informal) employment. In addition, the approach allows for any sign structure of covariances between the outcome equation residual ε and the various error terms (η) from the selection equation 10. Thus the regression of interest becomes:

$$log(w_{1,i}) = \beta_1 \mathbf{X}_i + \mu\left(P_1\right) + \nu_1$$

The function μ is approximated by a simple Taylor expansion which for simplicity it is assumed to be one. In practice, the procedure consist in two steps, (i) estimate the multinomial logit for three occupational choices, wage earner formal and informal and self employment and compute the predicted probability for each occupational choice. I don't consider other occupational choices such as unemployment or inactives because, I can observe only a set of instruments related to housing and household characteristics but cannot observer characteristics such as tenure, firm size or industry, thus I cannot compute properly the transitions probabilities. (ii) Include the predicted probability of the multinomial logt and assume sufficiency to include only the predicted probability of that particular occupational choice in the wage regression. Thus, the wage profile could be retrieved from a estimation of a regression separately for informal and formal workers following Topel (1991):

$$log(w_i^j) = \alpha + \sum_{k=1}^{3} \varphi_k tenure_i^k + \beta X_i + \mu\left(P_1^j\right) + \varepsilon_i^j, \quad j = formal, informal. \tag{12}$$

Where the term $\mu\left(P_1^j\right)$ is the predicted probability of choosing formal and informal estimated from the multinomial logit. The results are shown in Table 4 are the likelihood of choosing one occupation with respect to a base so Columns (1) describe the likelihood of choosing informal over formal employment. Notice that with higher tenure age, education and being male reduce the likelihood of choosing informal employment. Additionally, it is less likely to chose informality due to firm size, particularly for big firms. The instruments show that it is likely to choose informality over formality when the number of children are between one and three.

The opposite results are found for the likelihood of choosing formal employment over informal in Column (3). Its is likely to choose formality the higher the education, age, male gender tenure and firm size, particularly when firms are big. It is likely to choose working formally over informal in all industries except in manufacturing. Finally, notice the corrected wage regression in columns (5) and (6), the probabilities show that there is a negative selection that is statistically significant contrary to the probit however recall that this is the probability with respect to three occupational choices.

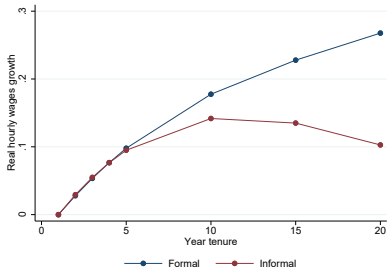
Table 4: Multinomial logit

	Logit				**OLS**	
Choice in the logit	**Informal**	**Self-emp.**	**Formal**	**Self-emp.**	**Informal**	**Formal**
Base outcome =	*Formal*		*Informal*			
	(1)	(2)	(3)	(4)	(5)	(6)
Tenure	-0.04***	-0.25***	0.25***	0.21***	0.04***	0.03***
Tenure^2	0.00***	0.01***	-0.01***	-0.01***	-0.00***	-0.00***
Tenure^3	-0.00***	-0.00***	0.00***	0.00***	0.00***	0.00***
Age	-0.03***	-0.05***	0.05***	0.02***	0.02***	0.02***
Age^2	0.00***	0.00***	-0.00***	-0.00	-0.00***	-0.00***
Education years	-0.01***	-0.07***	0.07***	0.06***	0.04***	0.08***
Gender (1=Male)	-0.25***	-0.32***	0.32***	0.07***	0.13***	0.09***
Firm size (No. employees)						
2 - 5	-5.40***	-0.11	0.11	-5.30***	0.30***	0.19***
6 - 20	-7.95***	-1.43***	1.43***	-6.52***	0.44***	0.45***
20 - 50	-8.72***	-2.47***	2.47***	-6.25***	0.44***	0.57***
50 >	-8.68***	-3.56***	3.56***	-5.12***	0.52***	0.69***
Industry						
Agriculture and Mining	-0.95***	-0.44***	0.44***	-0.51***	-0.02	-0.03
Manufacturing	-0.35***	0.22***	-0.22***	-0.56***	0.01	0.05***
Construction and retail	-0.61***	-0.50***	0.50***	-0.12***	-0.09***	-0.08***
Transportation and trade	-0.49***	-0.57***	0.57***	0.08**	-0.13***	-0.11***
Private Services	-0.43***	-0.39***	0.39***	-0.04	-0.14***	-0.10***
Instruments						
City	-0.02***	-0.03***	0.39***	0.30***		
Partner	0.00***	0.00***	0.11***	0.18***		
Number of dependants						
1 - 3	0.28***	0.36***	-0.36***	-0.08**		
4 - 5	-1.65	-13.11	13.11	11.46		
Home ownership	0.04**	-0.30***	0.30***	0.34***		
$\text{Prob}_{\text{Informal}}$					-0.21***	
$\text{Prob}_{\text{Formal}}$						-0.44***
Constant	7.18***	4.66***	-4.66***	2.51***	6.51***	6.25***
Observations	284,354	284,354	284,354	284,354	37,817	88,271
Pseudo R^2	0.56					
Log L	-121260.1					
LR (χ^2)	311857.6					
p-val	0.00					

Source: Statistics office and author's elaboration. *** p<0.01, ** p<0.05, * p<0.1.

Finally, the tenure wage profile in Figure 3 shows that by five years of tenure there is a wage growth gap of no gap by year 5 but grows to 16% by year 20.

Figure 3: Growth of hourly real wages and tenure with Multinomial Logit procedure



Source: Statistics office and author's elaboration.

4 Transitions into-out of formal informal employment

An important aspect in the wage growth and tenure relationship is that senior workers would suffer substantial wage losses if their jobs were to end. Thus, it is important to take into account that transitions could be negatively correlated to tenure. Many job transitions are the outcome of a career process by which workers are sorted into more durable and productive matches, thus high-wage jobs tend to survive and translates to long job tenures and higher wages, or alternatively, productive workers are keep by firms so they transition less often and will earn high wages.

In this matter, at any point in the worker's life time it is possible to transit between an informal and a formal job. Donovan et al. (2020) investigate if hiring (job-to-job transition) and separation (transition to employment) rates have a relationship with tenure and find a negative relation for low tenure but the relationship disappears as the tenure increases.[11] Namely, there is a negative strong trend in separation rates and job-to-job transition rates for workers who have been on the job for less than a year. By contrast, there is essentially no trend for workers who have five or more years of tenure; such workers in any country are unlikely to switch jobs or exit employment.

In the context of informal employment, the seminal work of Maloney (1999) there is evidence that workers transit between formal and informal jobs neglecting the duality of formal informal labor markets. This is the approach of Bosch and Maloney (2006), Bosch et al. (2007) and Bosch and Esteban-Pretel (2012), for the case of Mexico and Brazil where they find evidence that workers transit between formal and informal jobs using panel data. Contrary to those studies, the household survey data of Colombia is cross sectional, so it doesn't follow the individual for two or more periods. To address this issue, I use the retrospective questions of the survey that provide information on the current and past occupation state of the individual.[12] Although the questions permit to identify the current and previous occupational status the challenge comes from defining accurately informality in the previous job due to missing data.

[11] In Donovan et al. (2020), the employment-exit rate is constructed as the share of initially employed workers who transition to non-employment in the subsequent quarter. They also construct as one form of job transition the share of initially self-employed workers who transition to wage work.

[12] The relevant questions in the survey are depicted in the Annex 6.3

As in Bosch and Maloney (2006) it is assumed that the observed discrete-time mobility process is generated by a continuous-time homogeneous Markov process defined over a discrete state-space with finite possible states (occupational choice). Starting from the discrete tabulations, one can construct a discrete time transition matrix $p_{i,j} = Pr(t, t+n) = Pr(X(t+n)) = j|X(t) = i$ for $t = 0, ...$ and $n = 0, ...$ Where $p_{i,j}$ is the probability of moving from state i to state j in n steps. A maximum likelihood estimator of p_{ij} is:

$$p_{ij} = \frac{\overbrace{n_{ij}}^{\text{Transitions from i to j}}}{\underbrace{n_i}_{\text{Total number of observations initially in i}}}$$

Considering three states (formal wage earners, informal wage earners and unemployed) I compute transition probabilities from $t-1$ to t obtaining a three by three probability transition matrix.[13] Having defined the job transition of workers across all ages, I also compute the transitions dividing by age groups. The Table 5, shows that transition from formal to informal is U-shaped with a level of 2.1% for young workers and 3% for old. Conversely, informal to formal transition is decreasing across age groups and exhibit a low level compared to the former changing from 1.3% to 0.8%. In addition formal to formal transition are higher among young workers and older of around 3% but from informal to informal is less than 1%. This results complement other small literature on job flows and transitions in Colombia. López Castaño and Lasso Valderrama (2015) follow a similar approach but defining self employed as informal and find formal to informal transition probability of 4.5% and informal to formal 2.2%.

The results could depict that informal workers have a lower probability of transiting to either formality or another informal job compared to formal workers. In this case, there is a potential scarring effect of informality relative to formal workers.

Table 5: Discrete transitions probabilities by age groups, 2007-2019

	Age group								
	26-30	**31-35**	**36-40**	**41-45**	**46-50**	**51-55**	**56-60**	**61-65**	Average
Share formal wage earners	73.15	75.46	75.32	75.09	74.64	73.88	70.62	61.41	72.45
Share informal wage earners	26.85	24.54	24.68	24.91	25.36	26.12	29.38	38.59	27.55
Transition employment to employment									
Formal to formal	2.75	2.17	1.8	1.54	1.36	1.21	1.24	3.16	1.9
Formal to informal	2.14	1.72	1.55	1.45	1.52	1.71	2.02	3.11	1.9
Informal to informal	0.34	0.29	0.29	0.26	0.32	0.24	0.19	0.11	0.26
Informal to formal	1.3	1.07	0.98	0.96	0.92	0.98	0.87	0.78	0.98

Source: Statistics office and author's elaboration.

[13] I abstract from other states such as self employed and inactive to simplify the analysis.

5 Remarks for future research

Colombia is an example of a developing economy with a sizable share of informal employment (30%) that could be a potential source of major wage differential compared to formal workers. In this document, I explore if there is any informal formal wage gap related to tenure. Using two complementary empirical strategies, I consider two sources of selection bias into or out of wage earner occupation and into formal or informal employment to properly estimate a Topel (1991) regression. After controlling by worker and firms characteristics, both strategies point to a significant wage gap that grows nearly to 30% after 20 years. In addition, the transitions probabilities of informal and formal workers suggest that there is potential scarring effect for informal once their are in a formal job.

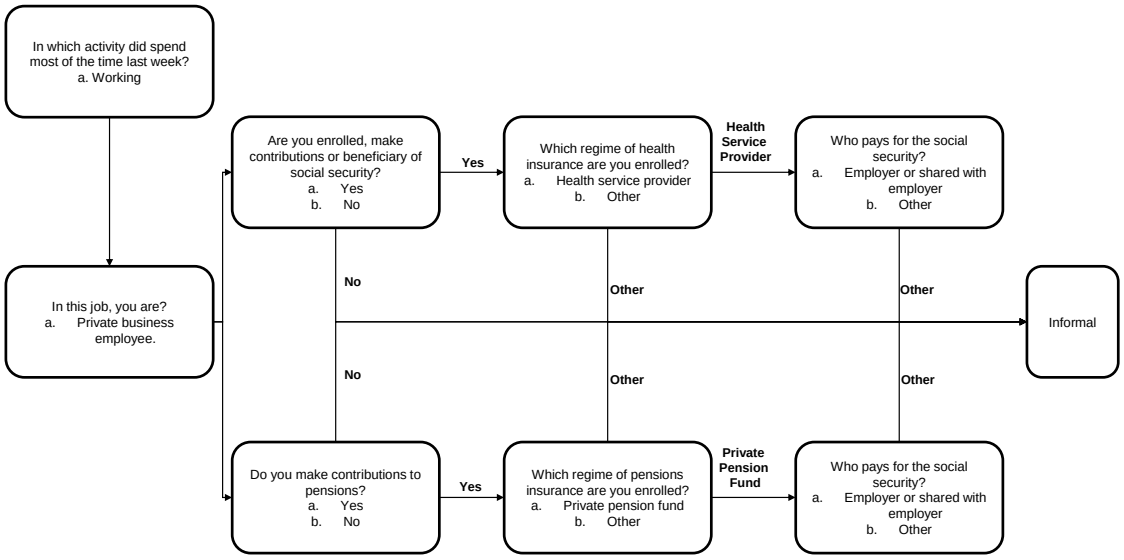
This finding could potentially be a starting point for future research to explain the underlying factors of this wage gap. First, it is important to explore the role of labor market institutions that are likely to explain informality persistence. On the workers side, informality persistence could be welfare policies and cash transfer programs, unemployment insurance, prevailing social safety net or regulation and tax avoidance. Bosch and Campos-Vazquez (2014), study a universal health coverage policy in Mexico and find that this was an incentive for firms and workers not to cover the cost of a formal contract and engage in informality. Likewise, cash transfer programs may have a have a similar effect because programs conditional on household characteristics, particularly, on unemployment status increases the incentive of workers to work informally but declare to be unemployed to remain eligible in the cash transfer program Bergolo and Cruces (2021).

Second, an additional path of research could be to study job training on the job training view, Ma et al. (2021) show that there are significant cross country differences on income because firms in developed economies invest more on job training. This could be an additional feature, informal workers receive less training than their formal counterparts and therefore experience low wage growth. This literature, has gain attention because of the increase of short term work arrangement as mention before. In addition working informally, implies there are less opportunities to having verified referrals for alternative jobs. This implies that, there is limited ex ante information for firms when hiring Beaman and Magruder (2012).

6 Appendix

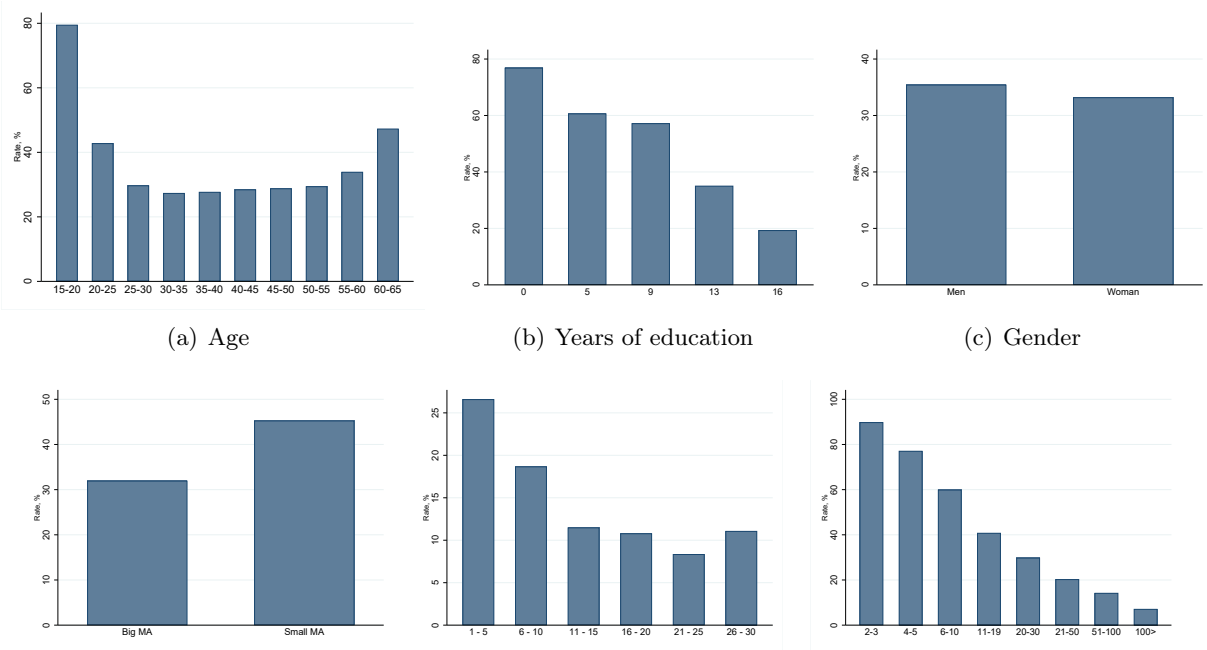
6.1 Survey questions to classify formal and informal employment

The survey includes questions on current employment and social security status to classify formal and informal workers questions to compute transitions. The relevant questions and answers are depicted in the following diagram:



Source: Statistics office and author's elaboration.

6.2 Demographic characteristics



(a) Age (b) Years of education (c) Gender

(d) City (e) Tenure (f) Firm size

Source: Statistics office and author's elaboration. Variables are computed with survey weights.

6.3 Survey questions for computing transitions

Occupational choice

- (QP6242) In which activity did spend most of the time last week?
  - a. Working, b. Searching for work, c. Studying, d. Home services, e. Permanent unable to work, f. Other.

If working the relevant questions are:

- (QP6426) For how long (in months) have you worked in this business/firm/office/farm/craft?
- (QP7020) Before your current job did you had another job? a. Yes, b. No.
- (QP7026) For how long (in months) did you work on your previous job?
- (QP760) For how long (in months) did you stopped working between your current and previous job?
- (QP7028) In your previous employment where you: a. Worker in private sector, b. Wage worker in public sector, c. Domestic employee, d. Self Employed, e. Owner or supervisor, f. Non paid family worker, g. Non wage worker in business, h. Day worker, i. Other.

If searching the relevant questions are:

- (QP7250) For how many weeks have you been searching?
- (Q7280) Have you search for job as a: a. Worker in private sector, b. Wage worker in public sector, c. Domestic employee, d. Self Employed, e. Owner or supervisor, f. Non paid family worker, g. Non wage worker in business, h. Day worker, i. Other.
- (QP7310) Have you search for a job for the first time or have you worked before at least for two weeks in a row? a. First time b. Have worked before.
- (QP7320) For how many week did you stopped working?
- (QP7350) In this last job you where a. Worker in private sector, b. Wage worker in public sector, c. Domestic employee, d. Self Employed, e. Owner or supervisor, f. Non paid family worker, g. Non wage worker in business, h. Day worker, i. Other.
- (QP7360) How many people in total had the firm/business/farm where you worked? a. Worked alone b. 2 to 3 c. 4 to 5 d. 6 to 10 e. 11 to 19 f. 20 to 30 g. 31 to 50 h. 51 to 100 i. 101 or more.

In addition, other questions are used to keep track on social security status and wage earners severance payments. Namely,

- (QP6140) Previously, where you enrolled, contributor or beneficiary of a health social security function? a. Yes, b. No.
- (QP6150) For how long (in months) you have not contributed or been enrolled in social security?
- (QP7460) Are you currently receiving unemployment subsidy? a. Yes, b. No.

6.4 Mapping of transitions with questions in the Survey data, 2007-2019

	State in t		
State in $t-1$	Formal wage earner	Informal wage earner	Search
Formal wage earner	1. Formal wage earner with tenure of a year or less. 2. Formal wage earner that had a previous formal job the year before the survey.	Informal wage earner with tenure of a year or less that had a formal job in the year previous to the survey.	Unemployed workers searching for a year or less and whose previous job was formal.
Informal wage earner	1. Formal wage earner with tenure of a year or less. 2. Formal wage earner that had a previous informal job the year before the survey.	Informal wage earner with tenure of a year or less that had an informal job in the year previous to the survey.	Unemployed workers searching for a year or less and whose previous job was informal.
Search	1. Formal wage earner with tenure of a year or less. 2. Workers searching for more than a year and less than two.	1. Informal wage earner with tenure of a year or less. 2. Workers searching for more than a year and less than two.	Unemployed workers searching for a year or more and first time workers searching for a job for more than a year.

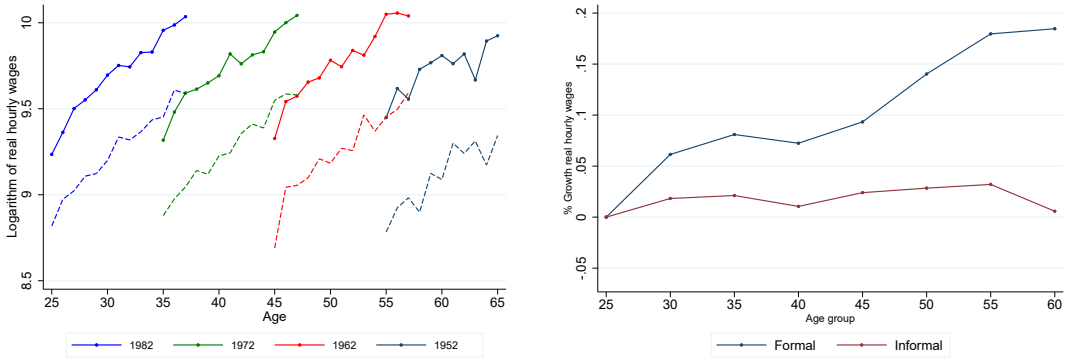
Source: Statistics office and author's elaboration.

6.5 Age-wage profiles

With regard to the broad picture of wage dynamics over the life cycle of workers, the data shows a significant difference of real hourly wage levels and growth between formal and informal workers. In Figure 6.1(a), each solid (dashed) line represents the average level of real hourly wage for some cohorts of formal (informal) workers, where each dot is the survey year, starting in 2007 at the bottom and ending in 2019 on the top of each line. Thus, the growth along a single line is the sum of time and age effects and the growth across lines are the cohort and age effects. The figure indicates two salient features: (i) the vertical formal-informal wage gap is positive, large and wider for older individuals which suggest time and age effects and (ii) the horizontal cohort gap remains stable for younger cohorts and changes slightly for older ones, suggesting relative equal cohort and age effects.

Moreover, in Figure 6.1(b) explore more the age affect assuming that individuals enter the labor market at age 25 as standard in the literature and compute the average wage across 5-year age bin and then the growth for each bin relative to the youngest age group.[14] The data depicts wage growth somewhat sustained for formal workers and stagnant for informal over their life-cycle which highlights a significant gap between formal and informal workers that reaches 20% by age sixty. This growth however, only depicts age effects and not the time effect discussed previously.

Figure 6.1: Hourly real wages by age-time-cohort decomposition, 2007-19, ages 25-65.



(a) Level log hourly real wages

(b) Growth log hourly real wages

Source: Statistics office and author's elaboration. Variables are computed with survey weights. Each solid line represent a specific cohort of formal workers and the dashed line the informal workers of the same cohort. Each dot in the solid line represents the reference survey starting at the bottom from 2007. Thus, the lines represent the average level of real hourly wages of that cohort (age group) from 2007 and 2019.

This observation relates to the extensive evidence of wage gap between formal and informal workers in developing countries but ads on how this gap is dynamic over the life-cycle of the workers; a feature often overlooked in the literature as mentioned in Ulyssea (2020).

A well-known difficulty in investigating how economic choices change over the life cycle is that it is impossible to separately identify the effects of age, time, and birth cohort on the outcome of interest, in a simple ordinary (OLS) framework. Note that cohort (birth year) is simply $cohort = time - age$ so this implies perfect collinearity. To solve this, Deaton (1997) proposes a simple way to achieve identification by normalizing the coefficients in the age-period-cohort specification. To fix ideas, consider the simplest

[14] Alternative, the approach could use potential experience of individuals to consider that each worker enters the labor market at different points in their life-cycle.

form of the model:

$$\log w_{it} = \Phi age_{it} + \gamma_t + \chi_t + \varepsilon_{it}$$

And consider an alternative form of writing the coefficients involved in the previous expression:

$$\tilde{\Phi}_t = \Phi + \kappa t \qquad \tilde{\gamma}_t = \gamma_t - \kappa t \qquad \tilde{\chi_c} = \chi_c - \kappa t$$

Where age_i indicates the ages of individual i, γ_t and χ_c are time and cohort dummies that indicate the year of the survey and the cohort group (5-year bin) and κ is any scalar. In practice, the chosen normalization is the one that is most suitable for the problem, attributing time-trends to the survey year, or to matching age and cohort effects. In this specification, where wages is the variable to be decomposed, a simple method is to attribute growth to age, and to use the year effects to capture cyclical fluctuations effects that average to zero over the long run.

A normalization that accomplishes this, makes the year effects orthogonal to a time-trend, so that $\mathbf{t}\gamma = 0$. This is a plausible case as Figure 6.1, describes that wage growth for different cohorts does no describe a clear pattern. The normalization can be achieved if wages are regressed on (i) dummies for group of cohort excluding the first, (ii) and a set of $T-2$ year dummies d_t defined from $t = 3, .., T$.

$$d_t^* = d_t - \Big[(t-1)d_2 - (t-2)d_1\Big]$$

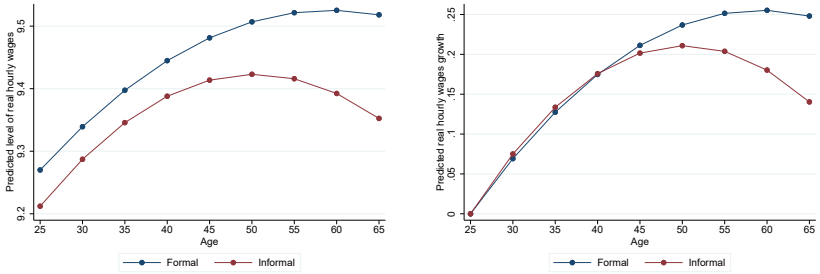
The underlying idea of this procedure is that if it is assumed that wages are constant over cohorts, ages, and years, the curves in Figure 6.1(a) collapse to a unique vertical line with slope 0. This vertical line could reflect a positive age effect each year compensated by an equal sized negative time effect, implying that, each cohort gets an age additional growth each year, but offset by a macroeconomic shock. In this context, younger cohorts would get less than older cohorts, because they are active in the labor market for much longer. However, as Figure 6.1(b) wages grow with age and stagnate or decrease at older ages. To capture this behaviour, it is useful to give each cohort a growth rate higher than the cohort born the previous year marking a negative trend in the cohort effect.

In practice, it is possible to recover an age profile controlling for time and cohort effects by regressing:

$$\log w_{it} = \alpha + \sum_{k=1}^{3} \Phi_k age_{it}^k \otimes I_{i,t} + \phi_1 I_{i,t} + \phi_2 age_{it} + d_t^* + \chi_c + \beta X_{it} + \varepsilon_{it} \tag{13}$$

The term age_i in the regression its specified as third degree polynomial interacted with a dummy $I_{i,t}$ that takes the value of one if the individual is a formal worker. In addition, X_t is a set of controls related to demographic characteristics (education years, gender, city, firm size, industry). Figure 6.2 shows the results of the wage-age profile both in levels and growth pointing to a significant differences in the level gap at each age point between formal and informal workers. Meanwhile, the growth rate of wages are similar up until age 40 and departs afterwards where stagnates for formals and decreases for informal workers. At the end of the life cycle the gap reaches 10%. Point estimates are depicted in Annex

Figure 6.2: Level and growth of hourly real wages and age



(a) Levels (b) Growth

Source: Statistics office and author's elaboration.

The previous result complements a small literature on wage gap between formal and informal workers for the case of Colombia. The estimated wage gap in the literature varies significantly from 30% to 60% because of the inclusion of other types of workers such as self employed, the definition of informality related to firm size and also the estimation techniques (Ramoni-Perazzi and Orlandoni-Merli, 2021). While this line of research documents the existence of a significant wage gap it does not speak about the growth over the life cycle.

Having discussed the estimated age profiles, it is important to mention that the identification procedure its subject to discussion. On the one hand, note that the parameter κ could be any scalar that could yield many potential results. As mentioned by Schulhofer-Wohl (2018) the standard method for solving this identification problem is to impose a normalization on the age, period, or cohort effects to pin down κ, but this depends on the arbitrary normalization used to pin down κ. If the normalization is incorrect, the estimator will be inconsistent.

References

Akcigit, U., Alp, H., and Peters, M. (2021). Lack of selection and limits to delegation: firm dynamics in developing countries. American Economic Review, 111(1):231–75.

Bargain, O. and Kwenda, P. (2014). The informal sector wage gap: New evidence using quantile estimations on panel data. Economic Development and Cultural Change, 63(1):117–153.

Beaman, L. and Magruder, J. (2012). Who gets the job referral? evidence from a social networks experiment. American Economic Review, 102(7):3574–93.

Becker, G. S. (1964). Human capital: A theoretical and empirical analysis, with special reference to education. University of Chicago press.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. Journal of political economy, 75(4, Part 1):352–365.

Bergolo, M. and Cruces, G. (2021). The anatomy of behavioral responses to social assistance when informal employment is high. Journal of Public Economics, 193:104313.

Bobba, M., Flabbi, L., and Levy, S. (2022). Labor market search, informality, and schooling investments. International Economic Review, 63(1):211–259.

Bobba, M., Flabbi, L., Levy, S., and Tejada, M. (2021). Labor market search, informality, and on-the-job human capital accumulation. Journal of Econometrics, 223(2):433–453.

Bosch, M. and Campos-Vazquez, R. M. (2014). The trade-offs of welfare policies in labor markets with informal jobs: The case of the "seguro popular" program in mexico. American Economic Journal: Economic Policy, 6(4):71–99.

Bosch, M. and Esteban-Pretel, J. (2012). Job creation and job destruction in the presence of informal markets. Journal of Development Economics, 98(2):270–286.

Bosch, M., Goni, E., and Maloney, W. F. (2007). The determinants of rising informality in brazil: Evidence from gross worker flows. World Bank Policy Research Working Paper, (4375).

Bosch, M. and Maloney, W. F. (2006). Gross worker flows in the presence of informal labor markets: the Mexican experience 1987-2002, volume 3883. World Bank Publications.

Bourguignon, F., Fournier, M., and Gurgand, M. (2007). Selection bias corrections based on the multinomial logit model: Monte carlo comparisons. Journal of Economic surveys, 21(1):174–205.

Bratti, M., Conti, M., and Sulis, G. (2021). Employment protection and firm-provided training in dual labour markets. Labour Economics, 69:101972.

Burdett, K. (1978). A theory of employee job search and quit rates. The American Economic Review, 68(1):212–220.

Burdett, K. and Mortensen, D. T. (1998). Wage differentials, employer size, and unemployment. International Economic Review, pages 257–273.

Cisneros-Acevedo, C. (2022). Unfolding trade effect in two margins of informality. the peruvian case. The World Bank Economic Review, 36(1):141–170.

Dahl, G. B. (2002). Mobility and the return to education: Testing a roy model with multiple markets. Econometrica, 70(6):2367–2420.

Deaton, A. (1997). The analysis of household surveys: a microeconometric approach to development policy. World Bank Publications.

Donovan, K., Lu, W. J., Schoellman, T., et al. (2020). Labor market dynamics and development. Federal Reserve Bank of Minneapolis, No. 596.

El Badaoui, E., Strobl, E., and Walsh, F. (2010). The formal sector wage premium and firm size. Journal of development economics, 91(1):37–47.

Eslava, M., Haltiwanger, J. C., and Pinzón, A. (2019). Job creation in colombia vs the us:âup or out dynamicsâ meets âthe life cycle of plantsâ. Technical report, National Bureau of Economic Research.

García, G. A. and Badillo, E. R. (2018). Rationing of formal sector jobs and informality: The colombian case. Journal of International Development, 30(5):760–789.

Garcia, I. L. et al. (2015). Human capital and labor informality in chile: a life-cycle approach. Unpublished manuscript, Department of Economics, University College London.

García-Pérez, J. I., Marinescu, I., and Vall Castello, J. (2019). Can fixed-term contracts put low skilled youth on a better career path? evidence from spain. The Economic Journal, 129(620):1693–1730.

Gasparini, L. and Tornarolli, L. (2009). Informalidad laboral en américa latina y el caribe: patrones y tendencias a partir de microdatos de encuestas de hogares. Revista Desarrollo y Sociedad, (63):13–81.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. (2017). Heterogeneous scarring effects of full-year nonemployment. American Economic Review, 107(5):369–73.

Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica: Journal of the econometric society, pages 153–161.

Heckman, J. J. and Li, X. (2004). Selection bias, comparative advantage and heterogeneous returns to education: Evidence from china in 2000. Pacific Economic Review, 9(3):155–171.

Jovanovic, B. (1979). Job matching and the theory of turnover. Journal of political economy, 87(5, Part 1):972–990.

Jovanovic, B. (1984). Matching, turnover, and unemployment. Journal of Political Economy, 92(1):108–122.

Lagakos, D., Moll, B., Porzio, T., Qian, N., and Schoellman, T. (2018). Life cycle wage growth across countries. Journal of Political Economy, 126(2):797–849.

Lazear, E. P. (1979). Why is there mandatory retirement? Journal of political economy, 87(6):1261–1284.

López Castaño, H. and Lasso Valderrama, F. (2015). Flujos de trabajadores en el mercado laboral colombiano, determinantes de las diferencias entre hombres y mujeres, y futuro laboral esperado. Technical report, IDB Working Paper Series.

Lucas, R. E. (2009). Ideas and growth. Economica, 76(301):1–19.

Ma, X., Nakab, A., Vidart, D., et al. (2021). Human capital investment and development: The role of on-the-job training. Technical report, University of Connecticut, Department of Economics.

Maloney, W. F. (1999). Does Informality Imply Segmentation in Urban Labor Markets? Evidence from Sectoral Transitions in Mexico. The World Bank Economic Review, 13(2):275–302.

Meghir, C., Narita, R., and Robin, J.-M. (2015). Wages and informality in developing countries. American Economic Review, 105(4):1509–46.

Ohnsorge, F. and Yu, S. (2021). The Long Shadow of Informality. World Bank, Washington, DC.

Perry, G. (2007). Informality: Exit and exclusion. World Bank, Washington, DC.

Quintin, E. (2008). Contract enforcement and the size of the informal economy. Economic Theory, 37(3):395–416.

Ramoni-Perazzi, J. and Orlandoni-Merli, G. (2021). Analysis of the formal/informal wage inequalities in colombia: A semiparametric approach. Journal of Applied Social Science, 15(1):107–131.

Salop, J. and Salop, S. (1976). Self-selection and turnover in the labor market. The Quarterly Journal of Economics, pages 619–627.

Schulhofer-Wohl, S. (2018). The age-time-cohort problem and the identification of structural parameters in life-cycle models. Quantitative Economics, 9(2):643–658.

Topel, R. (1991). Specific capital, mobility, and wages: Wages rise with job seniority. Journal of political Economy, 99(1):145–176.

Ulyssea, G. (2018). Firms, informality, and development: Theory and evidence from brazil. American Economic Review, 108(8):2015–47.

Ulyssea, G. (2020). Informality: Causes and consequences for development. Annual Review of Economics, 12:525–546.