Pontifícia Universidade Católica
do Rio de Janeiro

Eduardo Pinheiro Fraga

Selection on Ability and the Gender Wage Gap

DISSERTAÇÃO DE MESTRADO

Thesis presented to the Programa de Pós-Graduação em Economia of the Departamento de Economia, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Economia.

Advisor: Prof. Rodrigo Reis Soares
Co-advisor: Prof. Gustavo Mauricio Gonzaga

Rio de Janeiro
March 2014

Pontifícia Universidade Católica
do Rio de Janeiro

Eduardo Pinheiro Fraga

Selection on Ability and the Gender Wage Gap

Thesis presented to the Programa de Pós-Graduação em Economia of the Departamento de Economia do Centro de Ciências Sociais da PUC-Rio, as partial fulfilment of the requirements for the degree of Mestre.

Prof. Rodrigo Reis Soares
Advisor
EESP-FGV

Prof. Gustavo Mauricio Gonzaga
Co-advisor
Departamento de Economia - PUC-Rio

Prof. Claudio Abramovay Ferraz do Amaral
Departamento de Economia - PUC-Rio

Prof. Cecilia Machado
FGV/EPGE

Prof. Monica Herz
Coordinator of the Centro de Ciências Sociais - PUC-Rio

Rio de Janeiro, March 28th, 2014

Eduardo Pinheiro Fraga

The author graduated in Economics from Universidade de São Paulo – USP in 2012, he obtained the degree of master at PUC-Rio in 2014.

Bibliographic data

Fraga, Eduardo Pinheiro

Selection on Ability and the Gender Wage Gap/ Eduardo Pinheiro Fraga; advisor: Rodrigo Soares; co-advisor: Gustavo Gonzaga – 2014..

54 f. : il. ; 30 cm

Dissertação (Mestrado em Economia)– Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2014.

Inclui bibliografia

1. Economia – Teses. 2. Hiato salarial entre gêneros 3. Seleção diferencial 4. Seleção em não-observáveis I. Soares, Rodrigo Reis. II. Gonzaga, Gustavo Mauricio. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. IV. Título.

CDD:330

Acknowledgements

Rodrigo and Gustavo, for their excellent advising and for being my guides in the long process of creating this work,

Claudio and Cecilia, for their kindness to participate in the committee and for their valuable suggestions,

and Juliano, for helping me understand the strengths and weaknesses of my work.

Guilherme Hirata and Mauricio Fernandes, for their precious help with data.

The brightest class of which I have ever had the pleasure of being a part so far: my collegues in the master's and PhD programs at PUC-Rio, from all cohorts. Particularly, Josué, for helping me with Latex, and Tomás, for the countless discussions and for "his ideas".

My family and friends, who supported me when it was necessary and even when it was not really necessary.

All professors and staff at the Department of Economics of PUC-Rio.

And Gabriel de Abreu Madeira and Mauro Rodrigues Júniors, who first brought me into the world of academia.

Abstract

Eduardo Pinheiro Fraga; Soares, Rodrigo Reis (advisor); Gonzaga, Gustavo Mauricio (co-advisor). **Selection on Ability and the Gender Wage Gap.** Rio de Janeiro, 2014. 54p. Dissertação de Mestrado - Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

The literature generally emphasizes that female labor supply interruptions impact the life cycle evolution of the gender wage gap through reduced female work experience. We propose another mechanism: if women's selection on ability is different from men's, then interruptions would cause the gender gap to change over the life cycle. We use the RAIS dataset (a very large Brazilian employee-employer dataset) to assess this hypothesis, proceeding in two steps. First, we estimate Mincer equations controlling for worker fixed effects. Estimated fixed effects are then used as a proxy for ability in regressions in which the dependent variable is participation (various measures) and the explanatory variables are ability and its interaction with a gender dummy. Regression results suggest that selection on ability is more positive for men, providing an additional explanation for the early-career growth of the gender gap.

Keywords

gender wage gap; differential selection; selection on unobservables; ability; participation; RAIS

Resumo

Eduardo Pinheiro Fraga; Soares, Rodrigo Reis (orientador); Gonzaga, Gustavo Mauricio (co-orientador). **Seleção por Habilidade e o Hiato Salarial entre os Gêneros.** Rio de Janeiro, 2014. 54p. Dissertação de Mestrado - Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

A literatura mostra que interrupções na oferta de trabalho feminina impactam a evolução do hiato salarial entre os gêneros ao longo da vida por meio da redução na experiência feminina. Este trabalho propõe um mecanismo diferente: se a seleção por habilidade diferir entre os gêneros, então as interrupções causarão mudanças no hiato ao longo do ciclo de vida. Usamos a RAIS (um grande banco de dados brasileiro que conecta empregados a empregadores) para avaliar essa hipótese em duas etapas. Primeiro, estimamos equações Mincerianas controlando por efeitos fixos de trabalhador. Então, usamos os efeitos fixos estimados como proxies para habilidade em regressões nas quais a variável dependente é a participação (várias medidas) e as variáveis explicativas são a habilidade e sua intereação com uma dummy de gênero. Os resultados sugerem que a seleção por habilidade é mais positiva para os homens, o que explica parcialmente o crescimento do hiato salarial entre gêneros no começo da carreira.

Palavras-chave

hiato salarial entre gêneros; seleção diferencial; seleção em não-observáveis; habilidade; participação; RAIS

Sumário

List of figures

List of tables

1 Introduction

The gender difference in participation is one of the common explanations in the literature for the gender wage gap and its life cycle evolution. The high incidence of labor supply interruptions for women between ages 20 and 40 reduces their accumulated experience and wages in relation to men with similar characteristics (see Bertrand et al., 2010; Corcorant et al., 1993; Goldin and Katz, 2008). However, interruptions could affect the evolution of the gender gap through another mechanism, not yet fully appreciated by the literature: selection on unobservables.

Exit and entry in the labor market could affect the ability composition of the pool of women and men being compared to each other at each age. For example, a higher exit of skillful women as compared to their male counterparts would partially explain the growing wage gap between ages 20 and 40 (Fernandes, 2013). On the other hand, if selection is more positive for women than men, then the observed gap would underestimate the 'true' gap which would be obtained if non-working women and men were included in the calculation. In any case, differential selection by gender has potential implications for the life cycle evolution of the gap that have not been fully explored in the literature.

Among the various explanations for the life cycle evolution of the gender wage gap, many researchers have focused on the difference in participation between genders. From ages 20 to 40, women are more likely to take time-off or reduce their weekly working hours due to pregnancy and child rearing obligations. Thus, women's labor market experience tends to be smaller than their male counterparts'. However, most papers fail to take that difference into account. In most cases, work experience is proxied by 'potential experience' (time elapsed since leaving school). Mincer and Polachek (1974) were among the first to note that, while this strategy may be reasonably accurate for men, it overestimates women's experience. This in turn artificially increases the unexplained gender gap, since some of the female wage disadvantage that is due to their lower 'actual' experience is left unexplained.

Therefore, using better measures of experience could reduce the observed gap. For instance, Blau and Kahn (2011) build a measure of actual experience from the PSID. They show that substituting it for potential experience in the estimation of a Mincer

equation reduces the observed gap by up to 35%. Oaxaca and Regan (2009) and Fernandes (2013) find qualitatively similar results, although the reduction in the observed gap is much smaller for the latter, who uses Brazilian data.

Another approach to improve measures of experience is to focus on very specific groups of workers for whom precise work history data is available. For instance, Bertrand et al. (2010) use a database on MBA alumni from the University of Chicago. They show that the higher incidence of interruptions among women is one of the leading factors contributing to the wage gap. Corcorant et al. (1993) and Goldin and Katz (2008) perform similar analyses with graduates from Michigan Law School and Harvard (various fields), respectively. Both papers find that growing gender gaps can be at least partially explained by women's smaller full-time work experience.

Generally, the literature has linked participation and the life cycle evolution of the gender gap mainly through experience accumulation. However, we argue for another mechanism through which participation may affect the evolution of the gap: selection on unobservables, such as ability. If exit from the labor market is nonrandom, then it may affect the average characteristics of the pool of participating workers. For instance, if the less skilled are more likely to exit, then the pool's average skill will increase with time. Moreover, selection might differ by gender. If skilled women have relatively lower attachment to the labor market than skilled men, that could explain part of the growing gap in early working life: the 'best' women could be leaving faster than the 'best' men! The fact that changes in participation are large for women implies that this effect may be sizeable.

To empirically address this question, we use data on individuals' work history to recover a measure of ability and then analyze the relationship between this measure and labor market participation, separately by gender. Our data comes from the Brazilian RAIS (*Relação Annual de Informações Sociais*), a longitudinal employee-employer dataset covering the universe of Brazilian formal workers between 1995 and 2010. Each worker is identified by an ID number, allowing us to build a panel with their entire formal labor market history. We proceed in two steps. In the first step, we estimate a Mincer equation controlling for worker fixed effects (FEs), education, experience and other variables. Our experience variables are built directly from observable formal labor market history so they do not have the disadvantages of potential experience. Estimated FEs are interpreted as pecuniary measures of the set of unobserved abilities that are valued by the labor market, such as cognitive and non-cognitive skills (such as

commitment, motivation, etc). Then, in the second step, we estimate regressions in which the dependent variable is participation throughout 1995-2010 and the independent variables are the FE and its interaction with a dummy for the male gender ($male * FE$). We also control for education, birth cohort and (in some specifications) cumulative experience. The coefficient on the $male * FE$ variable measures the extent to which selection on ability (net of education) is stronger for males than for females. The implications for the life cycle evolution of the gender gap are straightforward.

Our main results suggest that selection on ability, albeit modest, is positive at early ages for both genders, and at later ages for the male gender. For instance, in our preferred specification, a one standard deviation increase in ability increases the probability of participation at age 25 in 7.5% for women and in 6.5% for men. More importantly, the coefficient on the $male * FE$ variable is positive and significant at all ages, implying that selection is more positive for men than for women. These results are qualitatively robust to controlling for different measures of previous 'actual' work experience. We hypothesize that gender differences in selection may be due to women's responsibility as child caregivers, to men's preference for 'up-or-out' careers, or to the combination of assortative mating with a negative effect of spousal income on female participation. We also find some evidence of a life cycle pattern in the strength of selection: the ability coefficient decreases after age 25 (in our preferred specification).

Generally, results suggest that men are more positively selected on ability than women. Therefore, the pool of working men improves more with time than the pool of working women, providing a new contributing factor to the life cycle evolution of the gender gap. In fact, a tentative quantitative analysis suggests that this mechanism could explain 39.5% of the gap growth between ages 21-36 after accounting for experience, education and other observable variables.

The main limitation of the RAIS dataset is that it only covers the formal labor market, leaving aside self-employment and the informal sector, which is sizable in Brazil. The omission of variables regarding experience in these activities could bias our estimates of selection if informal or self-employment experience correlates with ability while also affecting formal participation. We use data from the PNAD (*Pesquisa Nacional por Amostra de Domicílios*), the Brazilian annual household survey, to estimate these experience variables. We then include them as additional controls in our second step regressions, which does not change the main results.

Two papers in the literature are closely related to our work. Machado (2013) proposes and implements an IV-inspired estimator for the gender wage gap that is robust to arbitrary selection in the labor market. It even allows for the coexistence of negative and positive selection in different 'parts' of the market. But her work differs from ours in that it is silent about the nature of selection and it focuses on the secular tendency of the gap rather than on its evolution through the life cycle. Herrmann and Machado (2012) perform regressions of participation on cognitive ability (measured by tests) separately for men and women from four different cohorts. Even though their strategy resembles our 'second step', they use a direct measure of cognitive ability, while our measure (the FEs from a Mincer equation) potentially includes a wider array of characteristics such as non-cognitive ability. Moreover, they also focus on the secular evolution of selection (like Machado, 2013) rather than on its life cycle evolution. Other related papers are Blau and Kahn (2006) and Mulligan and Rubinstein (2008), who also consider the relationship between selection and the gender wage gap. But, once again, their focus is not on the life cycle evolution of the gap, but on its secular evolution.

The main contribution of this paper is to show that differential labor market participation of men and women affects the evolution of the estimated wage gap not only through human capital accumulation, but also through selection. Selection on ability may have important implications to the estimation of the gender wage gap over the life cycle. By recovering a comprehensive ability measure from Mincer equations and finding that its correlation to worker participation is higher for men than for women, we present evidence that high-ability men are more attached to the labor force than their female counterparts. This contributes to the increase of the gender wage gap as individuals age and skilled women exit disproportionately the labor market.

The rest of this paper is organized as follows. Section 2 presents a theoretical model by Cahuc and Zylberberg (2004) and uses it to discuss the literature and the idea of selection on unobservables. Section 3 presents our data and some stylized facts. Section 4 presents the methodology, explaining our two-step estimation procedure. Section 5 shows our main results. Section 6 provides robustness checks of the main results. Section 7 concludes.

2 Theoretical Background

In this section, we start by presenting a simple earnings determination model from Cahuc and Zylberberg (2004). We then use the model as a theoretical framework to discuss the findings in the literature and the effect of selection on unobservables on the estimation of the gender gap.

Consider a simple model, based on Mincer (1974), that extends the basic life-cycle model of human capital accumulation (Cahuc and Zylberberg, 2004). It allows workers to acquire human capital while employed, instead of focusing entirely on either working or studying at each moment.

Suppose an individual is born in period *0* and studies until age *t*, when she enters the labor market. Her working life ends at period *T*, when she retires. During 'schooling period' [0, *t*], her time is fully devoted to acquiring human capital. During working life (*t*, *T*], on the other hand, she can divide her time between training (which further increases her human capital stock) and working. For each instant $t + \tau$, let $s(\tau) \in [0, 1]$ be the fraction of time allocated to training, with the residual time being dedicated to work. Training increases the worker's stock of human capital, $h(\cdot)$, according to the following differential equation:

$$\dot{h}(t+\tau) = \rho_x s(\tau) h(t+\tau), \qquad \forall\, \tau \in [0, T-t] \qquad (1)$$

where ρ_x is 'the rate of return to training after leaving school'. Note that, the higher $s(\tau)$, the higher is the growth rate of human capital stock at instant $t + \tau$.

With competition in the labor market, the individual's income $y(\cdot)$ at each instant $t + \tau$ is given by the following equation:

$$y(t+\tau) = A[1 - s(\tau)]h(t+\tau), \qquad \forall\, \tau \in [0, T-t] \qquad (2)$$

where A is a productivity constant and $[1 - s(\tau)]$ is the fraction of time dedicated to work. From equation (2), we can see that an individual's earnings are proportional to her stock of human capital and to the fraction of the current period spent working. Therefore, it may be optimal to invest some time in acquiring human capital in order to increase future earnings potential, even if that means foregoing part of the current earnings potential.

Integrating equation (1) between $\tau = 0$ and $\tau = x$, we get $h(t+x) = h(t)e^{\rho_x \int_0^x s(\tau)d\tau}$. Substituting this equality into equation (2) yields:

$$y(t+x) = A[1-s(x)]h(t)e^{\rho_x \int_0^x s(\tau)d\tau}, \qquad \forall\, x \in [0, T-t] \qquad (3)$$

that is, the income of an individual with *x* years of experience depends on her stock of human capital upon graduation (*h(t)*) and on total time spent on additional training since graduation ($\int_0^x s(\tau)d\tau$). Mincer (1974) makes the simplifying assumption that the fraction of time spent on training (*s(x)*) declines linearly with *x*, the amount of time elapsed since graduation:

$$s(x) = s_0 - s_0\frac{x}{T}, \qquad \forall x \in [0, T-t] \qquad (4)$$

Taking logarithms on both sides of equation (3) and using equation (4):

$$ln\, y(t+x) = ln\, Ah(t) + ln[1-s(x)] + \rho_x s_0 x - \rho_x\left(\frac{s_0}{2T}\right)x^2, \quad \forall\, x \in [0, T-t] \quad (5)$$

Finally, applying the equality $h(t) = h(0)e^{t\rho_x}$ to equation (5):

$$ln\, y(t+x) = ln\, Ah(0) + t\rho_x + \rho_x s_0 x - \rho_x\left(\frac{s_0}{2T}\right)x^2 + ln[1-s(x)], \qquad \forall\, x \in [0, T-t] \qquad (6)$$

Thus, the model provides us with a classical theoretical motivation for estimating a Mincer equation, such as Equation (6). The log of earnings ($ln\, y(t+x)$) is a function of schooling (measured by *t*), work experience (terms *x* and x^2), 'hours worked' (*ln [1-s(x)]*) and the term *ln Ah(0)*, the log of the product of productivity and the initial stock of human capital with which the individual is 'born'. Since *A* and *h(0)* do not change with time and are positively associated with earnings, it seems reasonable to interpret *ln Ah(0)* as a proxy for the individual set of time-invariant 'abilities' that are valued by employers, such as cognitive and non-cognitive skills (e.g., motivation and commitment).

As explained in Section 1, economic literature has traditionally estimated equation (6) by proxying experience *x* and x^2 with potential experience (i.e. time elapsed since leaving school). The potential experience variable would often be calculated by using the formula: $x^P = age - educ - 6$, which implicitly assumes individuals start school at age 6 and work continuously after graduation. However, as pointed out by Mincer and Polachek (1974), the latter assumption is particularly inaccurate for women, who often take time off from their jobs due to family obligations. Thus, female potential experience x^P systematically overestimates their actual experience *x*. In other words, experience is *measured with error* and the error term correlates with gender. When

traditional papers ignored this measurement problem, they produced biased estimates of the gender gap. Intuitively, the difference in earnings due to women's lower experience could not be captured by the flawed potential experience variable, so it artificially inflated the estimated gender gap.

Recent authors have successfully addressed the problem of experience measurement by taking advantage of increasingly precise information on workers' work history (for instance: Bertrand et al., 2010; Blau and Kahn, 2011; Oaxaca and Regan 2009). By using this information, they greatly improve the precision of their measure of *x*, virtually eliminating the problem of measurement error. As expected, their results show that improving experience measures reduces the estimated gender gap.

It is clear from the discussion so far that literature has focused on experience *x*. However, we argue that the 'ability' term *ln Ah(0)* is also important because it may influence the gender wage gap in important ways. Specifically, the exit and entrance of workers in the labor force may be correlated to ability *ln Ah(0)*. In other words, workers may be *selected on ability*. Moreover, the strength (and even the sign) of the correlation may differ between the two genders, in which case it would impact the dynamics of the gender gap over the life cycle.

In order to better understand this argument, let us consider a slightly different, estimable version of equation (6). If we have panel data, we can estimate:

$$ln\, y_{it} = \omega_1 t_{it} + \omega_2 x_{it} + \omega_3 x_{it}^2 + \omega_4 hours_{it} + a_i + \Gamma_t + v_{it} \qquad (7)$$

where *y* is income, *t* is schooling and *x* is work experience (just like in equation 6), *hours* is weekly hours worked (which is analogous to $ln[1 - s(x)]$ in equation 6), a_i are worker fixed effects, Γ_t are time fixed effects and v_{it} is a random error term. If v_{it} is strictly exogenous (Wooldridge, 2010), then the estimation of (7) yields consistent estimates of equation (6)'s parameters (ρ_x, $\rho_x s_0$, $-\rho_x \left(\frac{s_0}{2T}\right)$). Also, estimated fixed effects $\hat{a}_i$ are consistent estimates for the 'ability' term $ln\, Ah(0)$.

Suppose that selection is positive and stronger for men than for women. Then the average of a_i for market participants will increase with age for both genders, but more so for men than for women. If one does not take workers' ability into account, it will be absorbed by the error v. Therefore, the average of v will grow for both genders over the life cycle, but faster for men, causing the unexplained gap to grow as well.

Of course, one could use the opposite argument: if female selection is *more positive*, then the 'true' gap would be even larger than the observed gap. The reason is that women who actually work would be too high in the ability distribution when

compared to male workers. Thus, whichever the case, selection on ability could have important effects on the gender gap. By estimating equation (7) and recovering the fixed effects estimates $\hat{a}_i$, we can analyze how much of the evolution of the estimated gender gap is due to differential selection over the life cycle.

3 Data

3.1. Sample and Construction of Variables

We use RAIS (*Relação Anual de Informações Sociais*), a dataset of administrative records collected by the Brazilian Ministry of Labor (MTE). Once a year, MTE requests that firms fill a form providing information on all employees who were formally employed in the firm at any moment of the previous year (Gerard and Gonzaga, 2013). Since all firms must send this information, RAIS covers the universe of the Brazilian formal labor market (including public employees). Each observation in the dataset consists of a contract-worker-establishment triplet in a specific year. Workers are identified by their PIS number (similar to a social security number), so they can be followed through different years and firms. Note that a worker can appear more than once in a given year, for instance if she worked for different firms or if she was fired and then hired again in that year.

The dataset includes: (i) firm-related variables, such as firm and establishment identifiers, sector of activity, size, state and municipality; (ii) worker-related variables, including the PIS number, gender, age and schooling; (iii) job-related variables such as the average real monthly earnings[1], occupation, contract weekly working hours, tenure, an indicator of whether the employment contract was still active on December 31st and, in case it was not, the reason and month of separation. If the worker was hired in that year, information about the month of hiring and the type of contract (e.g. temporary or permanent) is also provided.

The RAIS dataset is very large, with more than 55 million observations only in 2010. Since working with the full dataset is computationally impossible, we chose to work with a random sample. Our sampling algorithm works as follows. First, we build a list containing the PIS numbers of all workers born in 1974 who appear in the (full) RAIS dataset at some point between 1995 and 2010. [2] We collapse this list so that each

[1] The average of earnings is taken over all months of the year in which the contract was active. Nominal earnings of each month are deflated using the Brazilian consumer price index IPCA (*Índice de Preços ao Consumidor Amplo*).

[2] The advantage of using only one birth cohort is that our subsequent analyses will not be confounded by cohort effects. Moreover, individuals born in 1974 were relatively young (21 years old) in 1995, the first available year in our dataset. Thus, their (unobservable) working history prior to 1995 is unlikely to be either long or important from a human capital accumulation perspective.

worker's PIS number appears only once. Then, we take a random sample of PIS numbers (i.e. of workers) from this list. Following, we search for each of these workers in all available years (1995-2010). Thus, the resulting dataset contains the complete 1995-2010 work history of each sampled individual.

We perform some transformations on the 'raw' dataset in order to correct data inconsistencies regarding variables such as education and age. Details about these inconsistencies and our correcting procedures are available in Appendix A1. We also perform some minor additional data adjustments. We discard all observations with less than five or more than 60 weekly working hours and also observations with negative earnings, because these values are probably due to measurement error. We keep only the 'main job' that each individual holds in each year. We consider the main job to be the one with the highest average real monthly earnings. Finally, we discard individuals who appear in less than two years so that we can estimate the fixed effects through the Mincer equation for all workers in the dataset. The resulting dataset has 443,392 individuals, of which 44.1% are women. The total number of observations is 3,639,146.

We now briefly describe the additional variables that we build to use in our subsequent analysis. Our wage variable is *lwage*, the logarithm of average monthly earnings. Education variables are dummies *somecol* and *collegegrad*, indicating individuals with some college education (but who did not graduate) and college graduates, respectively. The base group includes all individuals with less than a college education. Birth year (*byear*) is computed by subtracting age from the current calendar year. We also generate sets of dummies for: current year (***year***), birth year (***byear***), age in 1995 (***dage95***), aggregated sector of firm activity (***aggsector***), firm size (***size***) and state (***state***).[3]

We build several experience variables. First, drawing from Spivey (2005), we have six 'nonlinear' experience vectors: ***actv***, ***FTactv***, ***FYactv***, ***MYactv***, ***FTFYactv***, and ***FTMYactv***. Each of these vectors has 15 dummies referring to each of the previous 15 years. For instance, vector ***actv*** contains 15 dummies (*actv1*, *actv2*, … , *actv15*), where $actvk_{itg}$ indicates whether individual *i* had any job in year $t - k$. The other five vectors are analogous to ***actv***, but refer to more specific types of job: full-time (FT), full-year

[3] ***aggsector*** comprises dummies for 26 broad sectors of firm activity, and ***state*** comprises dummies for the 26 Brazilian states, plus the Federal District. ***size*** comprises dummies for 10 categories of firm size, as measured by the number of employees.

(FY), most-year (MY), full-time full-year (FTFY), and full-time most-year[4] (FTMY) jobs, respectively. We also have a second group of experience variables: *prev*, *exper_prev*, *FTMYprev* and *experFTMY_prev*. Dummy *prev* indicates any job in the immediately previous year, and *exper_prev* is the total number of past years (excluding the immediately preceding year) in which the worker had any job. Variables *FTMYprev* and *experFTMY_prev* are analogous to *prev* and *exper_prev*, but refer to full-time most-year jobs (rather than any job).

We build two kinds of participation variables which will be used in the second step of our empirical analysis. First, there is variable *part*, a dummy for full-time most-year participation. Then, there are variables of the form $yearsFTMY_p$, which count the total number of years of period *p* worked full-time most-year. There are five such variables, referring to periods: 1995-1998, 1999-2002, 2003-2006, 2007-2010 and to the entire period 1995-2010.

3.2. Descriptive Analysis

Before moving on to the methodology section, we present some stylized facts and descriptive statistics. Since our focus is on the gender wage gap, it makes sense to start looking at it. In Figure 1, we plot the average logarithm of real earnings at each age, separately for men and women. We only include workers who were working full-time most-year at each age-gender. The figure shows that the gap has a life cycle pattern: it starts around 14.9 log points at age 21 and builds up to 20.6 log points at age 29 and 28.3 log points at age 36. This early-career growth in the gap echoes previous findings in the literature, such as in Li and Miller (2012), Bertrand et al. (2010) and Fernandes (2013).

For selection to have any influence on the evolution of the gender gap, there must be some entry and exit in the labor market for at least one of the genders, otherwise the pool of workers would be constant. To examine the extent to which women and men leave and enter the workforce, we plot the participation rate for each age-gender in Figure 2. To make the exercise more intuitive, the figure only includes workers who worked full-time most-year in our first sample year, 1995. Thus, participation can be

[4] Throughout this paper, we use full-time most-year participation as our measure of participation. We considered using Herrmann and Machado's (2012) full-time full-year definition of participation (at least 35 weekly hours and 50 annual weeks), but found it too strict for the Brazilian labor market, which is characterized by very high turnover (Corseuil et al., 2013). We define working 'most-year' as working for at least 9 months in a given year. We follow Herrmann and Machado (2102) in that full-time is defined as working at least 35 weekly hours.

interpreted as 'survival' in the formal labor market. The figure shows that exit is sizable for both genders, with the participation rate falling from 100%[5] at age 21 to around 42% at age 30 for females (52% for males). This opens up the possibility of changes in the composition of the pool of working men and women

Moreover, Figure 2 shows a surprisingly small difference between genders. Men's participation rate falls almost as quickly as women's. We believe that could be partially explained by the fact that men are more likely to leave formal occupations to either work in informal jobs or become self-employed. Since RAIS only covers formal workers, our data cannot distinguish between this kind of movement and exit from the market.

To assess this possibility, we use another database, the PME *(Pesquisa Mensal de Emprego),* conducted by IBGE (*Instituto Brasileiro de Geografia e Estatistica)*. The PME is a monthly urban labor force survey with a structure similar to that of the American CPS (Gerard and Gonzaga, 2013). Each household is interviewed monthly during two periods of four months, with an interval of eight months in between. The survey covers the six largest Brazilian metropolitan areas and it provides information about each household member (aged above 10) on variables such as occupation, formality and whether the individual is self-employed. The PME dataset[6] allows us to estimate (separately by age and gender) the probability of a worker transitioning to the informal sector or to self-employment, conditional on having left the formal sector.[7] Estimated probabilities are shown in Figure 3. It is clear that the probability of becoming informal or self-employed conditional on having left formality is consistently higher for men than for women. For instance: while male probabilities are always above 25%, female probabilities never reach 25%. This suggests that the difference in *formal* labor market exit shown in Figure 2 may actually underestimate the difference in *total* labor market[8] exit, since many men who leave formality are actually starting their own

[5] Note that participation at age 21 equals 100% by construction, since we excluded workers who did not work full-time most-year in 1995.

[6] We use the 'new' PME survey between years 2002 and 2010. We delete individuals with gender inconsistencies (i.e. who 'changed gender' across years).

[7] Estimation proceeds as follows: first, we keep only observations from the month of March in order to simplify calculations. For each age-gender, we compute the number of individuals who worked in the formal sector in a given year but did not work in the formal sector in the following year (l_{ag}). Then, we compute how many of them found a job in the informal sector (inf_{ag}) or became self-employed (se_{ag}) in the latter year. The (annual) probability of transitioning to informality or self-employment conditional on having left formality is then calculated for each age-gender ag as: $p_{ag} = \frac{\inf_{ag} + se_{ag}}{l_{ag}}$.

[8] We use the expression 'total labor market' in the sense of working in any kind of occupation that commands income: formal or informal employment, self-employment, etc.

businesses or finding a job in informality. Thus, they are not really leaving the workforce.

To check for robustness, we also repeat the exercise in Figure 2 for specific education groups. Panels A and B of Figure 4 present the results for individuals who never started college and for college graduates, respectively. Both graphs are very similar to Figure 2, the main difference being the higher level of participation of college graduates (Figure 4B) as compared to people without a college education (Figure 4A) or to the general population (Figure 2). In other words, educated workers are more attached to the (formal) labor force.

Panel A of Table 1 presents descriptive statistics for some variables in our dataset. Consistent with Figure 1, earnings and its logarithm have a higher mean for men than for women (1455.27 *vs* 1226.20 and 6.92 *vs* 6.76, respectively). Variance of wages is also larger for men, but we cannot infer whether this is because of higher wage inequality among men or because the male life cycle wage profile is steeper. Men constitute a larger portion (59.2%) of the observations in the dataset than women. Average age is slightly higher for females than for males (29.31 *vs* 28.92), which may suggest that women start working later than men. Average working hours, on the other hand, are higher for men (42.57 *vs* 40.54), which is consistent with the notion that women are more likely to work fewer hours (Bertrand et al., 2010).

Panel B of Table 1 presents the distribution of the education variable (separately by gender and for the sample as whole). Women seem to be relatively better educated than men, with a lower fraction of observations in the 'less than college' group (78.2% *vs* 88.7%) and a higher fraction in the 'college graduate' group (16.6% *vs* 7.9%). In any case, one sees that the vast majority of observations belong to the less educated group, for both genders.

4 Methodology

In this section, we describe our methodology. Our empirical analysis has two 'steps'. We first explain the methodology of our 'first step', in which we recover a measure of ability. Then, we present our 'second step', in which we investigate the relationship between participation and our measure of ability. The second step has two alternative strategies, so we explain each one separately.

4.1. First Step: Recovering the FEs

Our aim is to investigate selection on unobserved ability for women and men in the labor market. Therefore, the first step in our procedure recovers a measure of labor market ability, so that we can use that measure in the subsequent step. We do that by using our basic longitudinal dataset to estimate two Mincer equations, one for men and one for women, *controlling for worker fixed effects (FEs)* and other characteristics. After estimation, we can recover the fixed effects estimates. We interpret the FEs as pecuniary measures of the set of time-invariant unobserved abilities that are valued by employers: cognitive skills, commitment, motivation and 'soft skills', among others. The complete specification of the Mincer equation borrows from Fernandes (2013):

$$\begin{aligned} lwage_{itg} = {} & \boldsymbol{\delta_{1g}} \boldsymbol{exper_{itg}} + \boldsymbol{\delta_{2g}} \boldsymbol{educ_{itg}} + \beta_{1g} tenure_{itg} + \beta_{2g} hours_{itg} + \beta_{3g} age_{itg} \\ & + \beta_{4g} age_{itg}^2 + \boldsymbol{\alpha_{1g}} \boldsymbol{aggsector_{itg}} + \boldsymbol{\alpha_{2g}} \boldsymbol{size_{itg}} + \boldsymbol{\alpha_{3g}} \boldsymbol{state_{itg}} \\ & + \boldsymbol{\alpha_{4g}} \boldsymbol{year_t} + \gamma_{ig} + \varepsilon_{itg}, \\ & g \in \{f, m\} \end{aligned} \tag{8}$$

where g indexes gender (female or male), i indexes worker, and t indexes year. The dependent variable is the logarithm of earnings, *lwage*. Drawing from Spivey (2005), we control nonlinearly for experience: vector $\boldsymbol{exper_{itg}}$ comprises nonlinear control vectors ***actv***, ***FTactv***, ***FYactv***, ***MYactv***, ***FTFYactv***, and ***FTMYactv***, whose construction we explained in section 3.1 above. Education vector $\boldsymbol{educ_{itg}}$ comprises dummies for 'some college' and 'college graduate', respectively. Therefore, the omitted education dummy corresponds to people who had never started a college as of year *t*.

Variable $tenure_{itg}$ is the number of months elapsed (as of December of year *t*) since the hiring of individual *i* to her present job.[9] Variable $hours_{itg}$ is contract weekly working hours. Variables *age* and age^2 are the individual's age and its square. The other controls are sets of dummies for aggregated sector of activity (***aggsector***), firm size (***size***), state (***state***), and year (***year***). Finally, worker fixed effects are represented by γ, and ε is an error term assumed to be orthogonal to our explanatory variables.

The estimation of equation (8) (separately by gender) allows us to recover a measure of each worker *i*'s fixed effect γ_i. This measure of ability will be used in the second step of our analysis.

The main advantage of our specification for the Mincer equation is that it controls relatively precisely for past experience. As mentioned in Section 1, literature traditionally used potential experience as a proxy for actual experience, which distorts the estimation of the Mincer equation (Mincer and Polachek, 1974; Oaxaca and Regan, 2009). We, on the other hand, build our experience variables from each individual's work history as recorded in RAIS data, thereby improving the precision of the experience measure. This in turn allows us to estimate the Mincer equation more accurately, improving the precision of the fixed effects estimates.

However, there are also disadvantages of generating experience variables from RAIS data. First, data is silent about individuals' work history prior to 1995, the first year in our dataset. We believe this is not a severe problem, though, because individuals in our sample were very young in 1995 (21 years old) and are thus unlikely to have had much prior experience by then. Second, RAIS only covers formal jobs, leaving out informal and self-employment experience, which may be relevant. In the robustness section (Section 6), we try to address this issue by estimating a coarse measure of informal and self-employment experience and adding it to the 'second step' regressions as a control.

4.2. Second Step: Relationship between Ability and Participation

Estimation of the Mincer equation in the first step allows us to recover estimates of the fixed effects, $\hat{\gamma}$, which we interpret as measures of time-invariant labor market ability. We briefly discuss interpretation and normalization of $\hat{\gamma}$ and present some stylized facts about its distribution. Then, in our second step, we estimate participation

[9] If the job was terminated in year *t*, then the measure is taken in the month of separation, instead of December.

equations in which the main explanatory variables are $\hat{\gamma}$ and $male * \hat{\gamma}$, the interaction of measured ability with gender dummy *male*. We also control for education and other variables. The coefficient of $\hat{\gamma}$ in a regression is interpreted as measuring the sign and strength of female selection, while the coefficient of $male * \hat{\gamma}$ measures the gender difference in selection. We use two slightly different specifications for the second-step's participation regressions, which are explained in the two following subsections.

4.2.1. Strategy I

In the first approach (Strategy I, henceforth), we estimate one equation for each of three specific ages: 25, 30 and 35 years. The objective of these regressions is to show us a 'picture' of selection in different moments of the average woman's life cycle. Age 25 is still in the 'fertile phase' in which women have most of their children, while age 35 is probably past it, since most women have already had all of their children by that age. Thus, it is expected that selection might change across ages.

The equations for Strategy I can be written as:

$$\begin{aligned} part_{iga} = {} & \theta_{0a} + \theta_{1a}\hat{\gamma}_{ig} + \theta_{2a} male_{ig} * \hat{\gamma}_{ig} + \boldsymbol{\theta_{3a}} \boldsymbol{educ_{iga}} + \boldsymbol{\theta_{4a}} \boldsymbol{exper_{iga}} \\ & + \boldsymbol{\theta_{5a}} \boldsymbol{state_{iga}} + \theta_{6a} male_{ig} + \boldsymbol{\theta_{7a}} male_{ig} * \boldsymbol{educ_{iga}} \\ & + \boldsymbol{\theta_{8a}} male_{ig} * \boldsymbol{exper_{iga}} + \boldsymbol{\theta_{9a}} male_{ig} * \boldsymbol{state_{iga}} + \mu_{iga}, \\ & a \in \{25, 30, 35\} \end{aligned} \quad (9)$$

where g indexes gender, i indexes worker, and a indexes age. Participation variable $part_{iga}$ indicates whether individual i worked full-time most-year at age a. Dummy *male* indicates the male gender. Variable $\hat{\gamma}$ is the fixed effect obtained from the estimation of the Mincer equation, and $male * \hat{\gamma}$ is its interaction with *male*.[10] Vector $\boldsymbol{educ}$ contains education controls *somecol* and *collegegrad*, which also appear in equation (8). As in the Mincer equation, we control for state dummies (***state***). We also include interactions of *male* with all control variables: $male * \boldsymbol{educ}$ and $male * \boldsymbol{state}$. μ_{iga} is the error term.

First, we omit any experience controls. Then, we include experience variables so as to control for past on-the-job training (Herrmann and Machado, 2012): *prev* indicates whether the individual worked in any job in the immediately preceding year, and

[10] The $\hat{\gamma}$ variable included in participation equations is a normalization of the original FE. For each gender, we use the mean and standard deviation of that gender's FEs distribution. For more on normalization, see subsection 5.2.1.

exper_prev is the number of previous years (excluding the immediately preceding year) in which the individual worked in any job. Lastly, we substitute *prev* and *exper_prev* with their full-time most-year counterparts, *FTMYprev* and *experFTMY_prev*, thereby approximating the experience controls to the dependent variable. Note that experience variables control not only for accumulated training, but also for an 'inertial' aspect of participation. For instance, it is plausible that nonparticipation in the previous year decreases the probability of participation this year simply because it can be difficult to find a job. For each experience specification, we also control for the interaction of experience with the *male* dummy.

Our main coefficients of interest are θ_{1a} and θ_{2a}. The former is interpreted as the intensity of selection on ability for women (at age *a*), and the latter as the difference between men and women in the intensity of selection on ability (at age *a*). A positive θ_{1a} suggests that skilled women are more likely to be employed in formal occupations (positive selection), whereas a negative sign means that able women are more likely to leave the formal workforce (negative selection). For a given sign, a higher magnitude of θ_{1a} means that the 'strength' of female selection is higher. As for θ_{2a}, a positive value means that selection is stronger for men than for women. Since our focus is on the gender wage gap, θ_{2a} is the most relevant quantity for us. For example, if $\theta_{2a} > 0$, then selection is more positive for men, which will contribute to increase the gap over time because male 'stayers' are relatively more able than their female counterparts.

4.2.2. Strategy II

Our second approach (Strategy II, henceforth) is similar to the first one, but we estimate equations by time period, not by age. We split our total span (1995-2010) in four periods: 1995-1998, 1999-2002, 2003-2006, and 2007-2010. One equation is estimated for each time period. This strategy is equivalent to running regressions by age (such as in Strategy I) but with a wider 'window' of ages in each regression. Thus, Strategy II provides a clearer picture of different life phases but is silent about specific ages.

The equations for Strategy II can be written as:

$$yearsFTMY_{igp} = \rho_{0p} + \rho_{1p}\hat{\gamma}_{ig} + \rho_{2p} male_{ig} * \hat{\gamma}_{ig} + \boldsymbol{\rho_{3p}} \boldsymbol{educ_{ig}} + \rho_{4p} male_{ig} + \boldsymbol{\rho_{5p}} male_{ig} * \boldsymbol{educ_{ig}} + \vartheta_{igp}, \quad (10)$$

$$p \in \{95-98, 99-02, 03-06, 07-10, 95-10\}$$

where g indexes gender, i indexes worker, and p indexes time period. Note that, besides the four shorter periods, we also estimate equations for the entire period (1995-2010), yielding a total of five regressions.

Dependent variable $yearsFTMY_{igp}$ is the number of years of period p in which individual i worked full-time most-year. Therefore, its value is a number between zero and four (with the exception of $yearsFTMY_{ig,95-10}$, which lies between zero and 16). As in Strategy I, the main explanatory variables are $\hat{\gamma}$, the (normalized) fixed effect estimated from the Mincer equation, and $male * \hat{\gamma}$, its interaction with gender dummy *male*. Vector ***educ*** includes education dummies *somecol* and *collegegrad*, which also appear in equations (8) and (9). We also include the interaction of dummy *male* with vector ***educ***. Finally, ϑ_{igp} is an error term.

Similarly to Strategy I, our main interest lies on coefficients ρ_{1p} and ρ_{2p}. The same considerations regarding their interpretation and relevance for the gender wage gap also apply here.

5 Results

In this section, we present the results of our empirical analysis. We first show the results for the first step (Mincer equation) and discuss some characteristics of the fixed effects and normalization procedures. Then, we present the results for both strategies of the second step, which are our main results of interest. Finally, we present a tentative quantitative analysis of the results.

5.1. First Step – Mincer Equation

We start by briefly describing the results of our Mincer equations (first step). Table 2 presents the main coefficients. We omit the coefficients of year dummies, aggregated sector dummies, firm size dummies, state dummies and worker fixed effects. Since the coefficients of the experience variables are numerous, we also omit them from Table 2.

In Table 2, we see that estimated returns to education are slightly higher for men. The premium for completing college is 29.1% for females and 34.9% for males. There is also a nontrivial premium for people who start but do not finish college: 7.3% for women and 8% for men. The age profile of earnings (as reflected in the coefficients of *age* and age^2) is increasing and convex for all relevant ages for both genders.[11] Note that these age effects should not be interpreted as experience effects, but solely as seniority effects, because our Mincer equations control flexibly for experience.

The coefficient of *hours* is positive for both genders, as expected: working longer hours implies higher earnings. However, the return to working hours is considerably more positive for women (0.6%) than for men (0.36%). Finally, the *tenure* coefficient is negative for women and positive for men, although its magnitude is very small for both genders. This may be due to the fact that the regression controls for experience and age.

[11] Although the *age* coefficient is negative, the inclination of the age-earnings profile becomes positive at age 8.3 for men and at age 21.1 for women.

5.2. Fixed effects

The estimation of the Mincer equations allows us to recover the estimates of the worker fixed effects. In doing so, fixed effects become a new variable in our dataset that can be studied on its own. In this subsection, we discuss the interpretation and normalization of the fixed effects and present a brief descriptive analysis.

As mentioned above, fixed effects γ_{ig} in equation (8) are worker-specific time-invariant constants which add to workers' earnings in every period in which they are active in the labor market. We interpret γ as the monetary value of the set of personal 'inborn' abilities that are valued in the labor market, such as cognitive and non-cognitive skills. As for the unit of measurement, note that the dependent variable *lwage* in equation (8) is *the logarithm* of earnings, so γ (in its original form) should be read as a *percentage* increase in earnings owing to worker ability. For example, imagine two workers A and B with $\gamma_A = 0.3$ and $\gamma_B = 0.1$. If A and B had the exact same set of observable characteristics (such as education, experience, state, etc.), A would still command earnings 20% $(= 0.3 - 0.1)$ higher than B solely because of her higher ability. One should also remember that our Mincer equations are estimated separately by gender, so ordinal comparisons of γ are only valid within genders. For instance, it is correct to say that a woman with $\gamma = 0.5$ is 'more skillful' than another woman with $\gamma = 0.3$, but we cannot affirm that she is 'more skillful' than a man with $\gamma = 0.3$ because the values of γ were not generated from the same regression.[12]

In our 'second step' (Section 5.3) and robustness checks (Section 6), we use a normalized version of the estimated fixed effects, $\hat{\gamma}_n$, in lieu of the original $\hat{\gamma}$. The reason is that each of these empirical exercises divides our sample in subgroups (defined by gender), so we use normalization to 're-center' the FE distribution for each group. This makes the analysis more homogenous across groups, facilitating comparison. Formally, let g be a specific population group. Normalized fixed effects for this group are given by: $\hat{\gamma}_{ng} = \frac{\hat{\gamma} - m_g}{\sigma_g}$, where m_g and σ_g are the mean and standard deviation of the estimated FEs for individuals of group g.

One should note that normalization changes the interpretation of the magnitude of the fixed effects. They are now interpreted as the distance (in relevant standard

[12] In order to understand this point, it may be useful to use an extreme example. Suppose that women are much 'smarter' than men such that the most skilled man has lower ability than the less skilled woman. Then, there may be a man C with $\gamma_C = 1$ and a woman D with $\gamma_D = -1$, but we know that woman D is more skilled than man C by assumption.

deviations) between the individual's ability and the mean of the relevant FE distribution. Therefore, fixed effects no longer have a direct monetary interpretation. They measure the relative position of an individual in the ability distribution of the population group to which she belongs. Interpersonal comparisons of $\hat{\gamma}_n$ with the purpose of ranking ability are still only valid within genders, but inter-gender comparisons now have some meaning. For instance, if man A and woman B have the same value of $\hat{\gamma}_n$, then the percentile of the male ability distribution in which A is placed is approximately the same as the percentile of the female ability distribution in which B is.

We now briefly present some stylized facts about the distribution of the fixed effects. Figure 5 presents the FE distribution by gender. The mean is -0.045 for men and -.063 for women[13] and the standard deviation is 0.492 for men and 0.487 for women. Higher male variance can be inferred from the figure by noticing that the male distribution has 'fatter tails', that is, it has a higher concentration of individuals in the extreme values.

We also investigate the fixed effects distributions of specific population groups. Figure 6 presents distributions separately by educational attainment (that is, the highest level of schooling of each individual across all years). Generally, graphs in Figure 6 are similar to those of Figure 5: male distributions have higher means and variances than female distributions. Panels A and B of Figure 6 show FE distributions for individuals with no college education and for college graduates, respectively. Means are relatively high for college graduates (0.354 for women and 0.545 for men) as compared to those with no college education (-0.189 for women and -0.134 for men). This difference in skill between education levels is expected, since skilled individuals tend to attain higher levels of education. More interesting is the finding that the FE distribution of college graduates also has much larger *variance* than the distribution of individuals who never went to college: standard deviations are 0.62 and 0.67 for female and male graduates (respectively), which is almost twice as high as the standard deviations for the less-educated group (0.36 for females and 0.39 for males).

[13] Since FEs were estimated separately by gender, the fact that their mean is higher for men is uninformative. Moreover, note that the FE mean for each gender does not need to equal zero: Mincer equations and FEs are estimated on an unbalanced panel, but the FE distribution is computed over the cross-section.

5.3. Second Step

In this subsection, we present the results in the 'second step' of our analysis. Since there are two alternative empirical strategies in this second step, we present them separately.

5.3.1. Strategy I

In Strategy I, we regress a measure of full-time most-year participation on the normalized fixed effect (which is our proxy for ability, as explained above) and its interaction with gender dummy *male*. Table 3 presents results for each age (25, 30 and 35). Since the FE variable was estimated in a previous stage, we use bootstrap (50 repetitions) to estimate Table 3's standard errors.

Let us start by focusing on our simpler specification (columns 1, 4 and 7), in which we omit experience controls. Both the fixed effect coefficient and the $male * FE$ coefficients are positive and significant (at the 5% level) at all ages, suggesting the existence of positive selection in the labor market for both genders. For women, a one standard deviation increase in ability increases the probability of participation in 0.9–4.9 percentage points. For men, the corresponding magnitudes are 3.3–5.2 percentage points.[14] Moreover, positive and significant $male * FE$ coefficients at each age suggest that positive selection is stronger for men throughout the analyzed period. To get a better sense of the magnitudes of the coefficients, one can compare them with the full-time most-year participation rate for each age-gender. A one standard deviation increase in ability increases the participation rate between 2.3%–21.6%, depending on gender and age. For both genders, the impact of ability on participation decreases between ages 25 and 30 and decreases again between ages 30 and 35. This suggests a decreasing life cycle pattern in the importance of positive selection.

In our second specification (columns 2, 5 and 8), we include experience variables *prev* and *exper_prev* (for definitions, see subsections 3.1 or 4.2.1) and their interactions with *male* in order to control for previous investments in on-the-job human capital (Herrmann and Machado, 2012). The FE and $male * FE$ coefficients remain positive and statistically significant in all ages (with the exception of the FE coefficient at age 30). A one standard deviation increase in ability increases the probability of participation in 0.15–0.99 percentage points for women and in 1.3–2.3 percentage

[14] The effect of ability on male participation can be computed by adding the $male * FE$ coefficient to the FE coefficient, at each age.

points for men. Again, the $male * FE$ coefficient is positive and significant in all cases, implying that men's selection is stronger than women's. Note that the impact of ability is generally lower than in the first specification. A one standard deviation increase in ability now increases participation by only 0.5%–7.5%. Since experience controls are the only difference between specifications, it seems that the effect of ability on participation is partially mediated through experience. Skilled individuals are more likely to work at young ages, so they accumulate more human capital which either induces or helps them to keep working at older ages. Finally, the magnitude of selection now exhibits a U-shaped pattern, decreasing between ages 25 and 30 but increasing between ages 30 and 35.

The third specification (columns 3, 6 and 9), which is our preferred specification, substitutes experience controls *prev* and *exper_prev* (and their interactions with *male*) with their full-time most-year counterparts, *FTMYprev* and *experFTMY_prev* (and their interactions with *male*). This makes experience controls more similar to the dependent variable (full-time most-year participation). The FE coefficient is positive and significant at age 25, but not at ages 30 and 35. On the other hand, coefficients on $male * FE$ are positive and significant at all ages, once more implying stronger male positive selection. The impact of ability on participation is generally lower than in the second specification. A one standard deviation increase in ability raises participation in -0.1–1.7 percentage points for women, and in 0.5–2 percentage points for men, which correspond to modest -0.4%–7.5% increases. Additionally, the life cycle pattern in selection is similar to the first specification, with the magnitude of selection diminishing over time. Thus, it seems that positive selection is more relevant in earlier stages of the working life, and that its effects propagate throughout the life cycle chiefly through inertia and accumulation of human capital rather than by ability-related later-career exits and entrances.

The results of Strategy I imply that differential selection by gender may be one of the factors affecting the growth of the gender wage gap at early-career ages (see Figure 1). Since positive selection is stronger for men, male 'survivors' in the labor market tend to be selected from a higher portion of the ability distribution than female survivors. As the pool of male workers 'improves' faster than the pool of female workers, the gap increases. Of course, the magnitudes of the selection effect are modest, as shown in our third specification in Table 3, so it may be only one of various factors influencing the growth of the gap.

The results in this subsection may raise the question of *why* positive selection is stronger for men. Although we do not have a final answer, we hypothesize that men may have a preference for risky, 'up-or-out' careers which are more conducive to the 'survival of the fittest'. Another possibility is that the marriage market may influence female selection. In the presence of assortative mating, skillful women marry skillful men who tend to be wealthier and who can thus 'afford' to have a nonworking wife. Finally, the fact that mothers (rather than fathers) tend to assume the role of child caregivers may lead to female exit which is not necessarily related to ability. In any case, our data does not include information on marriage and children so it does not allow us to test these hypotheses. Further research in that direction using different datasets is warranted.

5.3.2. Strategy II

In Strategy II, we regress measures of participation in five different time periods on the normalized fixed effect and its interaction with gender dummy *male*. Results by time period are presented in Table 4. Since the FE variable was estimated in a previous stage, we use bootstrap (50 repetitions) to estimate Table 4's standard errors.

We start by looking at column 1, which shows the results for the entire 16-year time span (1995-2010). An increase of one standard deviation in an individual's ability increases her expected number of years worked full-time most-year in 0.41 and 0.69 (or 9.3% and 13%) for women and men, respectively. Both the FE and the $male * FE$ coefficients are statistically significant at a 1% level. Thus, as in Strategy I, it seems that positive selection is relevant to both genders, but more so to men than to women.

We now focus on the four specific 4-year periods (columns 2-5). Note that these four periods show us individuals in four specific 4-year age windows. Therefore, Table 4 allows us to examine selection in different phases of a person's life cycle. A one standard deviation increase in ability increases the expected number of years worked full-time most-year in 0.03-0.16 for women and in 0.15-0. 2 for men. In percentage terms, this is equivalent to 2.3%–20.5% increases in participation, depending on gender and time period (i.e. life phase). In each time period, both the FE and the $male * FE$ coefficients are positive and significant at the 1% level. Thus, it seems that selection is positive for both genders, and higher for men, in each and every life phase between ages 21 and 36, not only in the life cycle as a whole.

The main advantage of Strategy II over Strategy I is that each regression covers a wider window of ages, providing a clearer picture of broad life phases. Let us then

analyze what Table 4 tells us about life cycle patterns in selection. For women, we can see that selection is relatively strong at younger ages (columns 2 and 3), but it becomes weaker in the two later life phases (columns 3 and 4). Thus, there is evidence that the magnitude of female positive selection decreases with age. For men, the impact of ability on participation also exhibits a general downward trend, although there is a small increase between column 2 (ages 21-24) and column 3 (ages 25-28). Therefore, life cycle patterns on Table 4 echo those on Table 3 (third specification): positive selection is stronger in the beginning of workers' careers, but it decreases with age.

Overall, the results from Strategy II largely confirm those of Strategy I. Selection on ability stronger for men than for women. While the magnitude of selection may be modest, it does offer a partial explanation to the early-career growth of the gender wage gap.

5.4. Quantitative Analysis

Our main results suggest that stronger selection on ability for men contribute (albeit modestly) to the early-career growth of the gender wage gap. In this section, we perform a tentative quantitative analysis in order to get a clearer idea of the magnitude of this effect.

We estimate two Mincer equations, one controlling for fixed effects (FE model) and the other without FE controls (POLS model). Each regression uses our basic dataset and includes both men and women. Dependent and control variables are the same as in the first step's Mincer equation (equation 8), the only differences being that we remove age and its square and add age dummies (***age***), a dummy for the male gender (*male*) and the interaction of the two (***age****male*). Estimated ***age****male* coefficients show us the evolution of the residual gender wage gap (net of education, experience, and other controls) throughout the life cycle. Moreover, since the FE model controls for ability while the POLS model does not, comparison across models of the ***age****male* coefficients gives us an idea of the influence of selection on ability on the residual wage gap and its evolution. Table 5 shows the results.

Columns *iii* and *iv* show the estimated gender wage gap for the POLS and FE models, respectively. Note that, for the POLS model, it is necessary to add the coefficient on *male* (column *i*) to the coefficients on ***age****male* (column *ii*) in order to arrive at the gender gap. That is not the case for the FE model, in which the *male* coefficient cannot be estimated since gender does not vary with time. In both models,

the gender gap increases with age. Between ages 22 and 36, it grows from 10 to 30 log points in the POLS model, and from virtually zero to 13 log points in the FE model. For all ages, the FE model gap is smaller than the POLS model gap by 56%-97% (columns *v* and *vi*). In other words, when one estimates a Mincer equation without ability controls, at least half of the estimated residual gender wage gap can be explained by higher average ability among men. It is not that men are more skilled than women *in the population*, but rather that *working* men are selected from a higher portion of the ability distribution than *working* women. In other words, positive selection is stronger for men.

However, the main results in this paper point to the impact of selection on the life cycle *evolution* of the gender wage gap rather than on its size at any given age. Thus, one should ask how much of the *change* in the gap can be ascribed to selection. Since the only difference between the FE and POLS models is that the former controls for FEs, any excess growth of the POLS model gap over the FE model gap can be interpreted as being caused by differential selection on ability. With this in mind, we compute the annual variation of the POLS model gap (column *vii*) and the annual variation of the difference between the two gaps (column *viii*) and divide the latter by the former (column *ix*). Column *ix* shows that, on average, 39.5% of the annual growth of the POLS model residual gender gap can be explained by differential selection. Thus, it seems that a nontrivial portion of the early-career growth of the residual gap (after accounting for education, experience and other observables) owes to the higher exit of skilled women from the labor force as compared to skilled men.

6 Robustness: Controlling for Non-Formal Experience

One of the most important limitations of the RAIS dataset is that it only covers formal workers. Therefore, it does not allow us to observe workers' experience in self-employment and in the informal sector and to include it in our experience variables. Under certain circumstances, omitting 'non-formal' experience could bias the estimates of selection in our 'second step'. For instance, informal experience may be correlated to ability if less skilled individuals are more willing to accept informal jobs. It may also be correlated to full-time most-year formal participation, for example, if non-formal experience is a (imperfect) substitute for formal experience in the process of applying for a job. In this case, our regression estimates would underestimate selection *even when we control for formal experience*. Of course, non-formal experience could also correlate negatively with formal participation, for instance, if there is inertia in participation in the non-formal market. In this case, selection would be overestimated. Whichever the case, the omission of non-formal experience in our second step (Strategy I) regressions is potentially problematic.

To address this issue, we estimate workers' experience in self-employment and informality by using an auxiliary dataset, the PNAD (*Pesquisa Nacional por Amostra de Domicílios*), conducted by IBGE (*Instituto Brasileiro de Geografia e Estatística*). The PNAD is a nationally representative yearly household survey which covers all Brazilian states and provides information on variables such as occupation, formality and self-employment, among many others. PNAD data allows us to estimate rough measures of informal and self-employment accumulated experience (*exper_inf* and *expesr_self*, respectively).[15] We then include these measures as additional controls in the second and third specifications of our second step (Strategy I) regressions. We omit the first specification (no experience controls) because it does not make sense to control

[15] First, for each age-year-gender-state group *aygs*, we use PNAD data to compute the fraction of formally inactive individuals from that group who work in the informal labor market ($pinf_{aygs}$) and in self-employment ($pself_{aygs}$). Then, for each year when an individual *does not appear in the RAIS dataset* (i.e. when she does not have a formal job), we impute informal and self-employment participation by using her *aygs* group's $pinf_{aygs}$ and $pself_{aygs}$. For each individual-year, informal experience (*exper_inf*) and self-employment experience (*exper_self*) are the cumulative sums of imputed *pinf* and *pself* (respectively) in all previous years in which the individual did not appear in the RAIS dataset.

for non-formal experience when we are not even controlling for formal experience. Table 6 shows the results.

For both specifications, the FE coefficient is positive and statistically significant at ages 25 and 35, but not at age 30. On the other hand, the $male * FE$ coefficient is positive and significant for all ages. Thus, our main finding that selection is more positive for men than for women is robust to controlling for non-formal experience. As for the coefficient magnitudes, a one standard deviation increase in ability increases the participation rate in -0.2–1.1 percentage points (or -0.5%–4.5%) for women and in 0.4–2.3 percentage points (or 1.2%–7.6%) for men. For both specifications, the life cycle pattern in selection is similar to the one we found in the second specification of Table 3: the magnitude of positive selection declines between ages 25 and 30 and then rises between ages 30 and 35. Thus, it seems that the life cycle pattern suggested by our preferred third specification in Table 3 is only partially robust to including non-formal experience as an additional control.

Overall, we conclude that the analysis in this section echoes the main findings of our analysis in subsection 5.3.1 (namely, higher positive selection among men than among women) while presenting a different picture on secondary findings, such as the statistical significance of the FE coefficient at specific ages and the life cycle pattern of the magnitude of selection.

7 Conclusion

Interruptions in labor supply are more common for women than for men due to family reasons (such as pregnancy and child rearing). Therefore, much of the Labor literature focuses on differential accumulation of on-the-job human capital as an explanation for the early-career growth in the gender wage gap. However, we argue that interruptions also affect the evolution of the gap through another mechanism. Exit and reentry into the market are nonrandom. Particularly, they depend on unobservable worker ability. If selection on ability differs by gender, that could impact the evolution of the gap. For instance, if selection is more positive for men, then the pool of working men would improve faster than its female counterpart, causing the gap to increase with time.

We empirically test this proposition by using a two-step procedure on the Brazilian RAIS dataset. First, we use workers' history to recover a measure of individual ability. We do that by estimating Mincer equations controlling for worker fixed effects, which serve as a proxy for ability. Then, we perform regressions of participation variables on the estimated FEs (our measure of ability), also including other controls such as education. The FE coefficient in these regressions may be interpreted as indicating the sign and magnitude of selection on ability.

Our results suggest that positive selection on ability is indeed more relevant for men than for women. Therefore, male 'stayers' tend to be relatively more skilled than 'female' stayers, contributing to increase the gender wage gap over the life cycle. A tentative quantitative analysis suggests that a nontrivial 39.5% of the early-career growth in the residual gap (after accounting for observables such as education and experience) can be explained by this mechanism.

We also find some weak evidence of a life cycle pattern in selection. Particularly, our main analysis suggests that the direct impact of ability on labor supply decisions diminishes with time, at least between ages 25 and 35. However, since skillful workers accumulate more work experience in their early working life, they tend to remain in the workforce in their thirties. In other words, there is an inertial aspect to participation: ability influences participation in the mid twenties, and these decisions propagate throughout the thirties.

Our finding that male selection is more positive than female selection raises the question of why this is the case. We hypothesize that this may be explained by male preference for ‘up-or-out’ careers, by the disproportionate importance of women in childrearing, and by the combination of assortative mating with a negative influence of spousal income over female participation. However, the fact that the RAIS dataset lacks information on marriage and children limits our ability to test these hypotheses. Thus, we believe that using different datasets in order to evaluate these possibilities is a possible avenue for further research.

8 References

Bertrand, M., Goldin, C., and Katz, L. (2010). Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors. *American Economic Journal: Applied Economics 2 (July)*: 228–255.

Blau, F., and Kahn, L. (2013). The Feasibility and Importance of Adding Measures of Actual Experience to Cross-Sectional Data Collection. *Journal of Labor Economics 31(S1)*: S17 - S58.

Cahuc, P., and Zyberberg, A. (2004). *Labor Economics*. MIT Press Books, The MIT Books, edition 1, volume 1, chapter 4, pages 86-88.

Corcorant, M., Courant, P., and Wood, R. (1993). Pay Differences among the Highly Paid: The Male-Female Earnings Gap in Lawyers' Salaries. *Journal of Labor Economics 11 (July)*: 417-441.

Corseuil, C.H., Da Silva, A., Dias, R., and Maciente, A. (2010). *Consistência das Bases de Dados do Ministério do Trabalho, Relatório Final (Abril)* (Unpublished manuscript).

Corseuil, C.H., Foguel, M., Gonzaga, G., and Ribeiro, E. (2013). Youth Labor Market in Brazil Through the Lens of the Flow Approach. *Anais do XLI Encontro Nacional de Econometria*.

Fernandes, M. (2013). *Diferencial de salário por gênero: discriminação ou história profissional? Uma análise a partir dos dados da RAIS* (Unpublished doctoral thesis). Pontifical Catholic University of Rio de Janeiro, Brazil.

Gerard, F., and Gonzaga, G. (2013). *Informal Labor and the Cost of Social Programs: Evidence from 15 Years of Unemployment Insurance in Brazil.* Pontifical Catholic University of Rio de Janeiro (PUC-Rio) Discussion Paper No 608.

Goldin, C., and Katz, L. (2008). Transitions: Career and Family Life Cycles of the Educational Elite. *American Economic Review: Papers & Proceedings 98 (2)*: 363-369.

Herrmann, M. and Machado, C. (2012). Patterns of Selection in Labor Market Participation. *11th IZA/SOLE Transatlantic Meeting of Labor Economists*.

Li, I., and Miller, P. (2012). *Gender Discrimination in the Australian Graduate Labour Market*. Institute for the Study of Labor (IZA) Discussion Paper No 6595.

Machado, C. (2013). *Selection, Heterogeneity and the Gender Wage Gap* (Working paper). Retrieved from https://docs.google.com/file/d/0B-DH4T6okuytbFNmakhiZn A5UnM/edit.

Mincer, J. (1974). *Schooling, Experience and Earnings*, New York: National Bureau of Economic Research.

Mincer, J., and Polachek, S. (1974). Family Investments in Human Capital: Earnings of Women. *Journal of Political Economy 82 (March/April pt. 2)*: S76-S108.

Oaxaca, R., and Regan, T. (2009). Work Experience as a Source of Specification Error in Earnings Models: Implications for Gender Wage Decompositions. *Journal of Population Economics 22*: 463–499.

Spivey, C. (2005). Time off at What Price? The Effects of Career Interruptions on Earnings. *Industrial and Labor Relations Review 59 (October)*. 119-140.

Wooldridge, J. (2010). *Econometric Analysis of Cross-Section and Panel Data*. MIT Press Books, The MIT Press, edition 2, volume 1.

9 Appendix

Tables and Figures

Table 1: Descriptive Statistics

Panel A

	Women				Men				All			
Variable	**Mean**	**Stardard Deviation**	**Min**	**Max**	**Mean**	**Stardard Deviation**	**Min**	**Max**	**Mean**	**Stardard Deviation**	**Min**	**Max**
Earnings	1226.20	1737.51	54.79	78794.13	1455.27	2002.19	36.53	80416.27	1361.71	1901.88	36.53	80416.27
lw age	6.758	0.732	4.004	11.275	6.923	0.751	3.598	11.295	6.855	0.748	3.598	11.295
male	0	0	0	0	1	0	1	1	0.592	0.492	0	1
age	29.31	4.55	21	36	28.92	4.57	21	36	29.08	4.57	21	36
hours	40.54	6.86	5	56	42.57	4.19	5	60	41.74	5.53	5	60

Panel B: Education

	Women		Men		All	
Education level	**observations**	**%**	**observations**	**%**	**observations**	**%**
Less than college	1,162,368	78.20%	1,909,357	88.69%	3,071,725	84.41%
Some college	77,585	5.22%	74,426	3.46%	152,011	4.18%
College graduate	246,384	16.58%	169,026	7.85%	415,410	11.42%
Total number of observations	1,486,337		2,152,809		3,639,146	

Notes: variable *earnings* is average monthly earnings. Variable *lwage* is its natural logarithm. *male* is a dummy indicating that the individual is a man. Variable *hours* is contract weekly working hours, and *age* is the individual's age. All calculations use the full sample.

Table 2: First Step (Mincer Equation)

	Dependent variable: Logarithm of earnings	
	Women	Men
	(1)	(2)
somecol	0.0727***	0.0801***
	-0.00204	(0.00224)
collegegrad	0.291***	0.349***
	(0.00180)	(0.00213)
tenure	-4.79e-05***	0.000129***
	(1.61e-05)	(1.34e-05)
hours	0.00599***	0.00357***
	(8.04e-05)	(0.000108)
age	-0.00940***	-0.0789***
	(0.00170)	(0.00150)
age^2	0.000565***	0.00187***
	(2.96e-05)	(2.65e-05)
Observations	1,486,337	2,152,809
R-squared	0.294	0.305
Number of individuals	195,331	248,061

Notes: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1 . Both regressions use the full sample and control for experience variables, year dummies, state dummies, firm size dummies, aggregate sector dummies, and worker fixed effects.

Table 3: Second Step (Strategy I)

		Dependent variable: FTMY participation dummy								
		Age 25			Age 30			Age 35		
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
FE		0.0489***	0.00798***	0.0170***	0.0236***	0.00147	-0.00018	0.00856***	0.00993***	-0.0013
		(0.00145)	(0.00097)	(0.00098)	(0.00129)	(0.00108)	(0.00102)	(0.00130)	(0.00115)	(0.00095)
male*FE		0.00354**	0.0147***	0.00272**	0.0181***	0.0115***	0.00541***	0.0249***	0.00556***	0.00588***
		(0.00171)	(0.00133)	(0.00124)	(0.00170)	(0.00138)	(0.00128)	(0.00161)	(0.00151)	(0.00131)
Experience controls		X	*prev* and *exper_prev*	*FTMYprev* and *experFTMY_prev*	X	*prev* and *exper_prev*	*FTMYprev* and *experFTMY_prev*	X	*prev* and *exper_prev*	*FTMYprev* and *experFTMY_prev*
FTMY participation rate	Women	0.227	0.227	0.227	0.297	0.297	0.297	0.367	0.367	0.367
	Men	0.301	0.301	0.301	0.361	0.361	0.361	0.396	0.396	0.396
FE coefficient as % of participation rate	Women	21.6%	3.5%	7.5%	8.0%	0.5%	-0.1%	2.3%	2.7%	-0.4%
	Men	17.4%	7.5%	6.5%	11.5%	3.6%	1.4%	8.4%	3.9%	1.2%
Observations		443,392	443,392	443,392	443,392	443,392	443,392	443,392	443,392	443,392
R-squared		0.034	0.299	0.354	0.020	0.375	0.439	0.013	0.326	0.389

Notes: Bootstrapped standard errors (50 repetitions) in parentheses; *** p<0.01, ** p<0.05, * p<0.1 . Regressions use the full sample and control for the male dummy *male*, education dummies *somecol* and *collegegrad*, state dummies, and for the interaction of *male* with education dummies and state dummies. Additionally, regressions in columns 2, 5 and 8 control for experience variables *prev* and *exper_prev* and for the interaction of these two variables with *male*. Similarly, regressions in columns 3, 6 and 9 control for full-time most-year experience variables *FTMYprev* and *experFTMY_prev* and for their interaction with *male*. FTMY participation rate is the fraction of individuals working full-time most-year at each age-gender. FE is the normalized version of the worker fixed effects.

Table 4: Second Step (Strategy II)

		Dependent variable: number of years worked FTMY				
		1995-2010 (age 21-36)	1995-1998 (age 21-24)	1999-2002 (age 25-28)	2003-2006 (age 29-32)	2007-2010 (age 33-36)
		(1)	(2)	(3)	(4)	(5)
FE		0.408***	0.156***	0.149***	0.0694***	0.0326***
		(0.001262)	(0.000571)	(0.000606)	(0.000223)	(0.000105)
male*FE		0.287***	0.022***	0.0519***	0.0956***	0.117***
		(0.002063)	(0.000640)	(0.000660)	(0.000460)	(0.000515)
Average number of years worked FTMY	Women	4.407	0.762	0.989	1.221	1.435
	Men	5.335	1.009	1.281	1.463	1.583
FE Coefficient as % of average years worked	Women	9.3%	20.5%	15.1%	5.7%	2.3%
	Men	13.0%	17.6%	15.7%	11.3%	9.5%
Observations		443,392	443,392	443,392	443,392	443,392
R-squared		0.038	0.026	0.032	0.024	0.015

Notes: Bootstrapped standard errors (50 repetitions) in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Regressions use the full sample and control for male dummy *male*, education dummies *somecol* and *collegegrad*, and for the interaction of the education dummies with *male*. The table also reports the average number of years worked full-time most-year for each period-gender. FE is the normalized version of the worker fixed effects.

Table 5: The Impact of Controlling for FEs on the Estimated Residual Gender Wage Gap

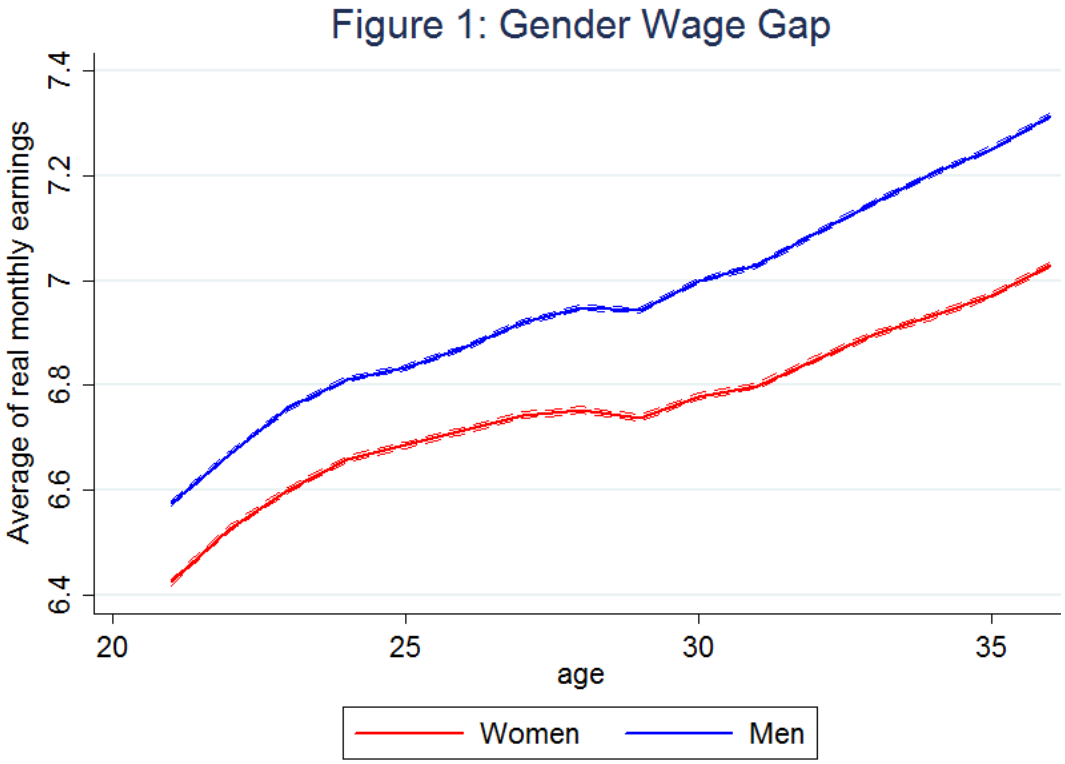
	POLS Model			FE Model					
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
Age	Coefficient on *male*	Coefficients on *age*male*	Gender wage gap (=i+ii)	Gender wage gap (Coefficients on *age*male*)	Difference between gaps (=iii-iv)	% difference between gaps (=v/iii)	Annual change in (iii)	Annual change in (v)	(viii) as a fraction of (vii)
-	0.295								
21		-0.198	0.097						
22		-0.191	0.104	0.0026	0.1015	97.5%	0.0070		
23		-0.174	0.121	0.0150	0.1061	87.6%	0.0170	0.0046	27.1%
24		-0.166	0.129	0.0163	0.1128	87.4%	0.0080	0.0067	83.8%
25		-0.153	0.142	0.0167	0.1254	88.3%	0.0130	0.0126	96.9%
26		-0.132	0.163	0.0320	0.1311	80.4%	0.0210	0.0057	27.1%
27		-0.116	0.179	0.0454	0.1337	74.7%	0.0160	0.0026	16.3%
28		-0.107	0.188	0.0481	0.1400	74.4%	0.0090	0.0063	70.0%
29		-0.0848	0.2102	0.0613	0.1490	70.9%	0.0222	0.0090	40.5%
30		-0.0658	0.2292	0.0785	0.1508	65.8%	0.0190	0.0018	9.5%
31		-0.058	0.237	0.0834	0.1537	64.8%	0.0078	0.0029	37.2%
32		-0.0491	0.2459	0.0883	0.1577	64.1%	0.0089	0.0040	44.9%
33		-0.0306	0.2644	0.1018	0.1627	61.5%	0.0185	0.0050	27.0%
34		-0.016	0.279	0.1156	0.1635	58.6%	0.0146	0.0008	5.5%
35		-0.0129	0.2821	0.1166	0.1656	58.7%	0.0031	0.0021	67.7%
36		0	0.295	0.1296	0.1655	56.1%	0.0129	-0.0001	-0.8%

Notes: Columns i, ii and iv show the coefficients of variables *male* and ***age***male* in a Mincer regression which also controls for education dummies, experience variables, tenure, weekly working hours, year dummies, state dummies, aggregate sector dummies, firm size dummies, a set of age dummies ***age*** and, in column iv, worker fixed effects. *male* is a dummy for the male gender and ***age***male* is the interaction of this dummy with the set of age dummies ***age***. In column iii, we calculate the residual wage gap at each age for the model without FEs (POLS model) by adding the *male* coefficient to the ***age***male* coefficients. The residual wage gap at each age for the model with FEs (FE model) is equal to the ***age***male* coefficients in column iv. Column v shows the difference between gender wage gaps calculated in columns iii and iv, and column vi shows this difference as a fraction of the gender wage calculated in column iii. Column vii shows the annual variation in the gender wage gap calculated in column iii, and column viii shows the annual variation in column v. Column ix shows the latter variation as a fraction of the former.

Table 6: Second Step (Strategy I) with Controls for Non-Formal Experience

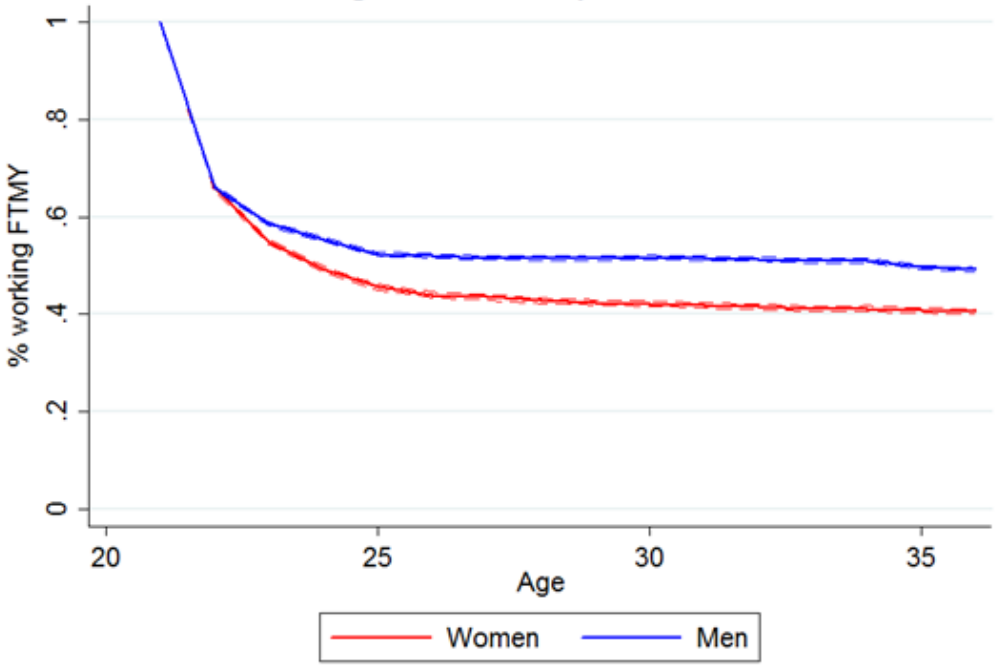
		Dependent variable: FTMY participation dummy					
		Age 25		Age 30		Age 35	
		(1)	(2)	(3)	(4)	(5)	(6)
FE		0.00801***	0.0103***	0.00166	-0.00158	0.0109***	0.00247***
		(0.00097)	(0.00097)	(0.00108)	(0.00100)	(0.00114)	(0.00093)
male*FE		0.0149***	0.00746***	0.0114***	0.0058***	0.00486***	0.00284**
		(0.00133)	(0.00124)	(0.00138)	(0.00126)	(0.00150)	(0.00133)
exper_inf		-0.0508**	-0.100***	-0.0714***	0.0777***	0.0114	0.198***
		(0.02263)	(0.01396)	(0.01269)	(0.00581)	(0.00957)	(0.00469)
exper_self		-0.0263	-0.492***	-0.056***	-0.269***	-0.0847***	-0.222***
		(0.03092)	(0.02868)	(0.01110)	(0.00914)	(0.00736)	(0.00592)
male*exper_inf		0.0391*	0.153***	0.0413***	-0.0362***	-0.0171	-0.145***
		(0.02206)	(0.01377)	(0.01333)	(0.00620)	(0.01084)	(0.00514)
male*exper_self		-0.0346	0.216***	-0.0218*	0.143***	0.0351***	0.142***
		(0.03483)	(0.02748)	(0.01137)	(0.00986)	(0.00815)	(0.00605)
Experience controls		*prev* and *exper_prev*	*FTMYprev* and *experFTMY_prev*	*prev* and *exper_prev*	*FTMYprev* and *experFTMY_pre*	*prev* and *exper_prev*	*FTMYprev* and *experFTMY_pre*
FTMY participation rate	Women	0.227	0.227	0.297	0.297	0.367	0.367
	Men	0.302	0.302	0.362	0.362	0.396	0.396
FE coefficient as % of participation rate	Women	3.5%	4.5%	0.6%	-0.5%	3.0%	0.7%
	Men	7.6%	5.9%	3.6%	1.2%	4.0%	1.3%
Observations		442,954	442,954	442,954	442,954	442,954	442,954
R-squared		0.299	0.364	0.375	0.446	0.327	0.395

Notes: Bootstrapped standard errors (50 repetitions) in parentheses; *** p<0.01, ** p<0.05, * p<0.1. Regressions use the full sample with the exception of individuals for whom state-gender-age-year information on the probability of informality or self-employment is missing for at least one year in which they were not working formally. Regressions control for the male dummy *male*, education dummies *somecol* and *collegegrad*, state dummies, and for the interaction of *male* with education dummies and state dummies. Additionally, regressions in columns 1, 3 and 5 control for experience variables *prev* and *exper_prev* and for the interaction of these two variables with *male*. Similarly, regressions in columns 2, 4 and 6 control for full-time most-year experience variables *FTMYprev* and *experFTMY_prev* and for their interaction with *male*. FTMY participation rate is the fraction of individuals working full-time most-year at each age-gender in the sample used in the regressions. FE is the normalized version of the worker fixed effects.



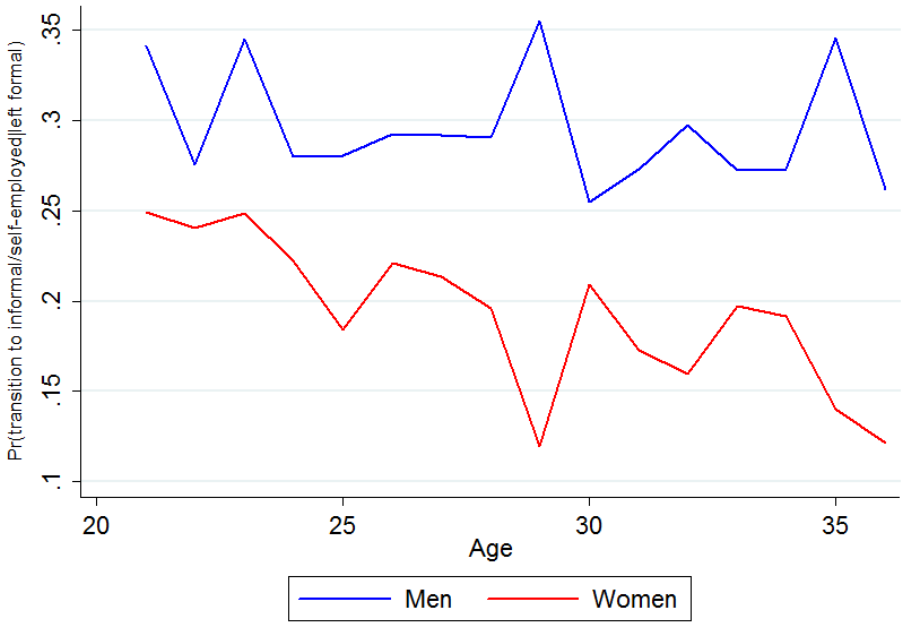
Notes: For each age-gender, the figure shows the average of *lwage* (the logarithm of earnings). The full sample is used. For each age-gender, computation of the average only includes individuals of that gender who were working full-time most-year at that age. Dashed lines show 95% confidence intervals.

Figure 2: Participation Rate

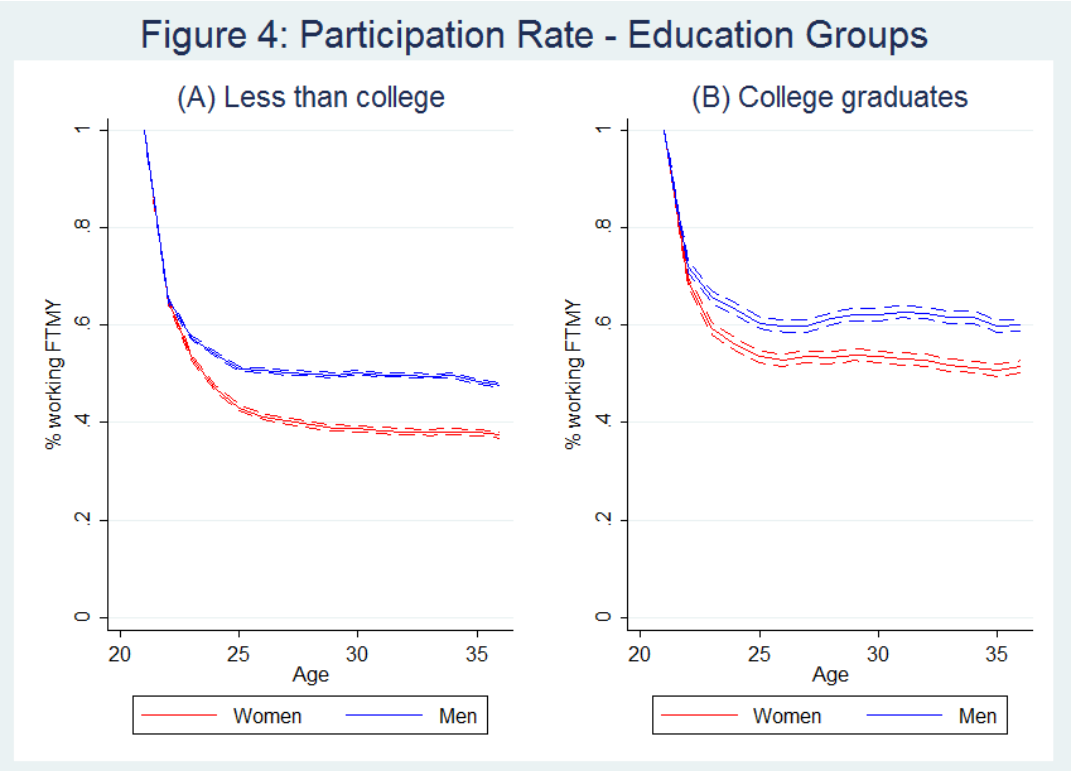


Notes: For each age-gender, the figure shows the percentage of individuals of that gender who were working full-time most-year at that age. Only individuals who worked full-time most-year in 1995 are used in the computation. Dashed lines show 95% confidence intervals.

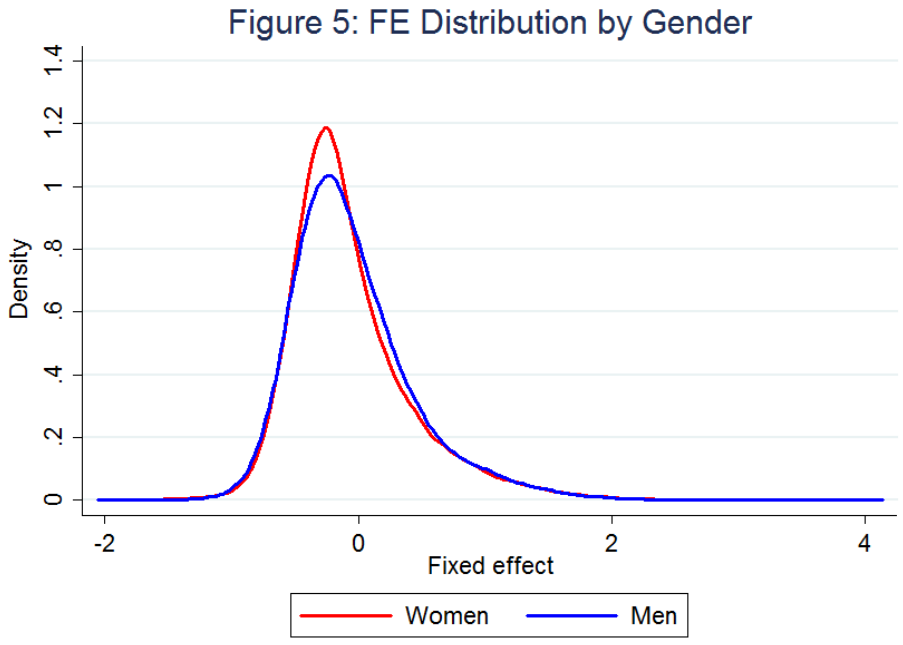
Figure 3: Probability of Transition to Informal/Self-Employed Conditional on Leaving Formality



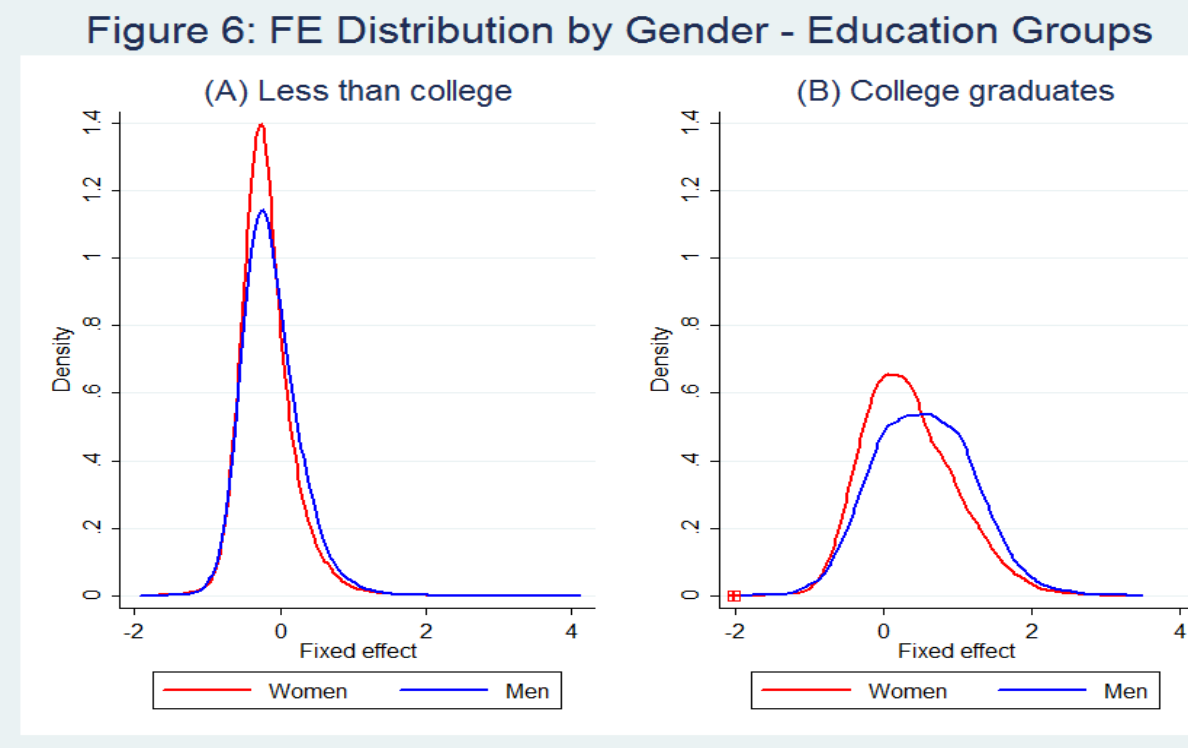
Notes: For each age-gender, the figure shows the probability that an individual of that age- gender will work in informality or in self-employment in year *t+1*, conditional on the fact that she worked in a formal job in year *t* and did not work in a formal job in year *t+1*. Probabilities were estimated from the PME (*Pesquisa Mensal de Emprego*) dataset. Only observations from years 2002-2010 and from the month of March were included in the calculation.



Notes: For each age-gender, the figure shows the percentage of individuals of that gender who were working full-time most-year at that age. Only individuals who worked full-time most-year in 1995 are used in the computation. Panels A and B only include in the computation individuals whose terminal level of education was 'never started college' and 'college graduate' (respectively). Dashed lines show 95% confidence intervals.



Notes: For each gender, the figure shows the estimated density function of the estimated fixed effects for individuals of that gender. Kernel density estimation uses the Epanechnikov kernel. The full sample is used.



Notes: For each gender, the figure shows the estimated density function of the estimated fixed effects for individuals of that gender. Kernel density estimation uses the Epanechnikov kernel. Panels A and B only include in the computation individuals whose terminal level of education was 'never started college' and 'college graduate' (respectively).

9.1. Data Inconsistencies and Correcting Procedures

In this section, we describe the data inconsistencies in the original database and the procedures we use to correct these problems, when possible.

Some individuals have inconsistent age information (e.g., some of them age three years in one year). For each observation, we compute the implied birth year by subtracting age from the current year. For individuals with two different and adjacent implied birth years (e.g. 1973 and 1974), we assume that the correct birth year is the one that appears more often, and we recalculate the individual's age in each year accordingly. If the two different birth years are not adjacent (e.g. 1973 and 1975), we assume that the correct birth year is the one that appears in at least 75% of observations, recalculating age accordingly. If no birth year has a frequency of at least 75%, the individual is deleted. Individuals with more than two different implied birth years are also discarded.

Note that we act less conservatively when the two implied birth years are adjacent than when they are not. The reason is that, in the former case, age information is not necessarily inconsistent. For instance, suppose a worker born in June 1974 is fired from his job in March 2000 and then hired and fired again in October 2001. Age equals 25 in his 2000 entry (his age upon being fired for the first time) and 27 in his 2001 entry (his age upon being fired for the second time). Thus, this worker will have two different implied birth years (1975=2000-25 and 1974=2001-27), even though there is no inconsistency in his age information.

There are also individuals with inconsistent gender information, that is, they 'change gender' at least once. Part of these errors is due to the fact that MTE imputes the male gender to observations with invalid gender information (Corseuil et al., 2010). Of course, part of the errors may also come from other sources of measurement error. Therefore, it is unclear how we could try to correct these inconsistencies. Since accurate gender information is crucial for our strategy, we chose to be conservative, deleting all workers with inconsistent gender information.

Some observations appear to be missing from the original dataset. For instance, some individuals worked for a firm in t and $t + 2$ but do not appear in that firm in $t + 1$, even though the data does not show either separation in t or hiring in $t + 2$. In cases like this, in which there is only one 'missing year', we artificially create a $t + 1$ observation. Working hours and earnings are linearly interpolated using adjacent values. We use an analogous procedure for cases in which there are two 'missing years', that is, an individual worked in a firm in t and $t + 3$, but she does not appear in that firm in $t + 1$ or $t + 2$, even though the data does not show either separation in t or hiring in $t + 3$. For cases in which there are three or more 'missing years', the individual is deleted.

Many individuals have inconsistencies in the education variable, that is, their education decreases over time (e.g. an individual who appears as a college graduate in year 2000 but as a high school graduate in year 2001). We use the algorithm developed by Fernandes (2013) in order to correct these inconsistencies whenever possible. Where there is a 'drop' in education, the algorithm essentially uses the adjacent values to impute a more 'reasonable' value either in the year in which the drop occurred or in the year prior to the drop. For example, if there are many years in which education equals 'high school' with only one year of 'college graduate' in the middle, the algorithm changes the latter value to 'high school'. Not all education inconsistencies can be reasonably corrected, so the resulting education variable is missing for some workers.

Since education is an important control in our subsequent analysis, these workers are discarded.

For some observations, the state where the firm is located is missing. Since state is a control variable in the subsequent analysis, all workers for who state information is missing in *some* year are deleted from the dataset.

As mentioned above in Section 3.1, we also: keep only individuals born in 1974; delete all observations with negative earnings; delete all observations with less than five or more than 60 weekly working hours; keep only the 'main job' (i.e. the job with highest earnings) for each individual-year; and discard all workers who appear in the dataset in only one year.

Table A1 summarizes all data procedures and shows the number of remaining individuals (by gender) at each stage of the transformations. The final dataset (line *ix*) contains 443,392 individuals, of which 195,331 (44.1%) are women. Note that the correction of the education variable is not possible for many individuals, decreasing the size of the sample by 18.7% (lines *vi*-*vii*). Also, data inconsistencies seem to be more common for men, as implied by the fact that the percentage of women increases with data correction procedures.

Table A1 - Data Transformation and Sample Size

Procedure number	Description of procedure	Number of remaining individuals				Percentage of women
		Women	Men	Inconsistent gender information	Total	
	Original dataset	282,682	413,419	45,394	741,495	38.1%
(i)	Correct inconsistent age information + delete individual when not possible	276,802	393,379	39,937	710,118	39.8%
(ii)	Keep only individuals born 1974	269,666	376,052	36,023	681,741	39.6%
(iii)	Delete individuals with inconsistent gender information	269,666	376,052	0	645,718	41.8%
(iv)	Correct missing observations + delete individual when not possible	268,106	373,652	0	641,758	41.8%
(v)	Delete observations with hours < 5, hours > 60, or earnings < 0	267,005	372,410	0	639,415	41.8%
(vi)	Keep only the 'main job' at each year	267,005	372,410	0	639,415	41.8%
(vii)	Correct errors in education + delete individual when not possible	232,755	286,910	0	519,665	44.8%
(viii)	Drop individuals with missing state information	232,645	286,823	0	519,468	44.8%
(ix)	Delete individuals who appear in only one year	195,331	248,061	0	443,392	44.1%