COLUMBIA | SIPA
Center for Development Economics and Policy

CDEP-CGEG WORKING PAPER SERIES

CDEP-CGEG WP No. 50

Selection on Ability and the Early Career Growth in the Gender Wage Gap

Eduardo Fraga, Gustavo Gonzaga, and Rodrigo R. Soares

February 2018

COLUMBIA | SIPA
Center on Global Economic Governance

Selection on Ability and the Early Career Growth in the Gender Wage Gap[*]

Eduardo Fraga[♣]

Gustavo Gonzaga[†]

Rodrigo R. Soares[‡]

September 2017

Abstract

This paper analyzes the effect of selection on ability on the evolution of the gender wage gap during the first years of professional life. We use longitudinal data with 16 years of the early career history of formal sector workers in Brazil. The panel allows us to build a measure of unobserved ability that we use to analyze the dynamics of labor market selection across genders as individuals age. We focus on the cohort born in 1974, for which we have a close to complete history of formal labor market participation. For this cohort, the average ability of formally employed men improved in relation to that of women during the first years of professional life. The selection of men and women into the labor market was similar at age 21, but by age 31 high-ability men (one standard deviation above the mean) had a probability of employment 1.6 percentage point higher than their high-ability female counterparts. This contributed to the increase in the conditional gender wage gap observed in the early career, as the ability distribution of employed women deteriorated in relation to that of employed men. Our estimates suggest that, for the 1974 cohort, this mechanism explains 32% of the cumulative growth in the conditional gender wage gap between ages 21 and 36.

Keywords: gender wage gap, selection, ability, lifecycle
JEL Codes: J16, J21, J31, J71

[*] The authors wish to thank Mauricio Fernandes and Guilherme Hirata for invaluable help with the data. This paper benefited from comments from Joseph Altonji, Claudio Ferraz, Tomas Guanziroli, Judith Hellerstein, Joseph Hotz, Lawrence Kahn, Cecilia Machado, Naercio Menezes-Filho, Duncan Thomas, and seminar participants at Duke University, Insper, PUC-Rio, University of Connecticut, University of Maryland at College Park, Yale University, the 4th SOLE/EALE World Meetings (Montreal, 2015), the 1st International REAP Meetings (São Paulo, 2015), and the 4th LACEA Labor Network Annual Meeting (Cartagena, 2016).

[♣] Yale University; *eduardo.fraga* at *yale.edu*

[†] PUC-Rio; *gonzaga* at *econ.puc-rio.br*

[‡] Columbia University and Sao Paulo School of Economics-FGV; *r.soares* at *columbia.edu*

1. Introduction

Labor market participation evolves differentially across genders over the lifecycle, with women typically displaying lower labor force attachment starting in the early twenties. The implications of this pattern for the accumulation of experience and market-specific human capital have been highlighted in the literature as one of the main reasons behind the widely documented increase in the gender wage gap during the early professional life (see, for example, Corcorant et al., 1993; Goldin and Katz, 2008; Bertrand et al., 2010). But differential labor market participation across genders can also affect the evolution of the gender wage gap through another channel, not yet fully appreciated by the literature: selection on unobserved ability. If the groups of women and men being compared to each other at each age do not have the same underlying distribution of abilities, then the age-specific gender wage gap may partially reflect differential changes in the composition of the labor force.

In particular, if high-ability men have a relatively higher probability of staying – maybe due to positive assortative mating and to the effect of spousal income on female labor force participation – then part of the early career growth in the gender wage gap would be explained by differential change in participation across genders. On the other hand, if high-ability women are more attached to the labor market and have a relatively lower exit rate over the lifecycle – maybe due to higher wages – then the glass ceiling effect would be more important than suggested by the early career growth in the gender wage gap. The implications of differential gender selection for the evolution of the wage gap are not theoretically obvious and, at the same time, have not received enough attention in the empirical literature.

This paper tackles this question by using a panel with 16 years of the early career history of formal sector workers in Brazil. The panel allows us to build a measure of unobserved ability that we use to analyze the dynamics of labor market selection across genders and its evolution during the early stages of the professional life. Our panel is constructed from the Brazilian official registry of workers (from now on, RAIS, from *Relação Anual de Informações Sociais*), a longitudinal employee dataset collected by the Brazilian Ministry of Labor containing information on the universe of formal workers in Brazil from 1995 to 2010. We focus on a single cohort and on individuals for whom we have close to complete descriptions of formal labor market history, and therefore restrict the sample to individuals born in 1974 (who were 21 years old in 1995).

We start by conducting a very simple descriptive exercise that highlights the importance of selection for the early career growth in the gender wage gap. Using our RAIS sample, we estimate a traditional wage equation allowing for a non-parametric age profile of the gender wage gap (interactions of gender and age dummies). We show that the age profile of the gender wage gap estimated from the RAIS dataset is close to that observed for employees in commonly used representative household surveys from Brazil (such as PNAD, the Brazilian National Household Survey). In our RAIS sample from the 1974 cohort, the conditional gender wage gap starts at 13% at age 21 and grows monotonically to 32% by age 36. Following, we re-estimate the wage equation in the RAIS dataset using individual fixed effects. Though we cannot estimate the level of the gender wage gap under this specification, we can still recover its variation across ages (through the interactions of gender and age dummies). We then compare the evolution of the gender wage gap during the first 15 years of this cohort's professional life across the two specifications, with the understanding that the fixed effects account for unobserved variation in productivity across individuals. The results of this exercise indicate that, for the 1974 cohort, selection explains 32% of the cumulative growth in the conditional gender wage gap between ages 21 and 36, and more than 50% before age 30.

Our main empirical exercise takes advantage of the panel structure of the RAIS dataset to describe in detail how the relationship between ability and formal employment evolves during the early career. In order to achieve this goal, we proceed in two steps. First, we construct a measure of unobserved ability from the individual-level fixed effects of wage equations, controlling for demographics and for a detailed characterization of past labor market history. To make sure that discrimination itself does not contaminate our estimates of individual unobserved ability, we estimate separate wage equations for each gender and normalize the estimates of fixed effects. Our comparisons of ability across genders are based on this normalized transformation of individual fixed effects, which gives a measure of the set of unobserved skills – both cognitive and non-cognitive – valued by the labor market. The direct comparison of this ability variable across genders is valid under the assumption that the distribution of skills for individuals who participate in the formal labor market at some point between 1995 and 2010 is the same across genders. Otherwise, absent this hypothesis, this comparison is still informative about the relative changes in the ability distribution across genders.

Using this measure of ability, we analyze how the pool of young men and women born in 1974 participating in the formal labor market changed between ages 21 and 36. First, we show graphically that the ability distribution of employed women deteriorated in relation to that of men between ages 21 and 36. Following, we analyze the pattern of labor market selection through time by running regressions where the dependent variable indicates employment at different moments between ages 21 and 36, and the independent variables of interest are our measure of individual ability and an interaction of ability with a gender dummy (male). The interaction of ability with the gender dummy indicates whether selection on ability was stronger or weaker for men as compared to women. The pattern of this coefficient across ages captures the evolution of this differential selection during the early career of this cohort of workers.

Our main result suggests that, for the 1974 cohort, selection on ability was positive for both genders during the first years in the formal labor market, but with age became relatively more important for males. A one standard deviation increase in ability led to an increase in the probability of employment at age 21 of 4.6 percentage points for women and 5 percentage points for men. But while overall selection on ability was reduced over time, its difference across genders increased: by age 31, the effect of a one standard deviation increase in ability on the probability of employment was 1.2 percentage point for women, and 2.7 percentage points for men. This differential effect was somewhat reduced by age 36, but still remained large and statistically significant.

We explicitly address the main threats to our empirical strategy. First, our results rely on estimates of individual ability that are conditional on an extensive set of controls capturing a detailed characterization of previous work experience in the formal sector, and are also robust to the inclusion of additional controls capturing potential experience in the informal sector. So our measure of ability does not reflect unobserved labor market experience. Second, results are robust to specifications that account for the possibility of selection on time-varying unobservables, which in principle could bias the estimates of individual ability in ways that might be correlated with gender (given that the elasticity of labor force participation is thought to be larger for women than for men). Third, our main results are not sensitive to particular methodological choices motivated by concerns about the precision of the ability estimates (choice of weights in the regression), the comparability of ability distributions across genders (normalization procedure), and the possibility of different returns to productive attributes across genders (running a single regression *versus* gender-specific regressions in our "first step"). Finally, our main result remains valid when we re-estimate the entire procedure separately by level of schooling (though the specific profile of evolution in differential selection across genders varies by level of schooling).

Trying to identify the determinants of the dynamics of selection of men and women into the formal labor market is beyond the scope of this paper. But it is not difficult to conjecture that the documented patterns may be associated with the timing of fertility decisions, as the evidence provided by Bertrand et al. (2010) and Adda et al. (2017) would suggest, or with the related issue of higher female demand for flexibility in work schedules (as in Goldin, 2014; Goldin et al., 2017), which in the context of a developing country may be linked to positions outside the formal sector.[1]

A large literature has analyzed the implications of differential labor market entry and exit across genders for the evolution of the gender wage gap with a focus on accumulation of experience rather than on selection on unobserved ability. Mincer and Polachek (1974) were among the first to notice that the use of "potential experience" (age minus schooling minus six) as a proxy for actual experience artificially increases the growth in the gender wage gap, since potential experience overestimates actual experience in a way that is correlated with gender. Various papers have since used improved measures of actual labor market experience and more comparable samples of men and women to show that part of the gender wage gap – and of its early growth – can be attributed to lower actual labor market experience among women (Corcorant et al., 1993; Goldin and Katz, 2008; Oaxaca and Regan, 2009; Bertrand et al., 2010; Blau and Kahn, 2013; Fernandes, 2013).

Though related to the point discussed here, the focus of this literature is different from ours. Among these authors, Bertrand et al. (2010) are the only ones who hint at the idea that differential change in selection may also be relevant. They show, using a highly selected sample (University of Chicago MBAs), that women who marry and have children – and, therefore, are more likely to leave the labor force – are, if anything, positively selected in terms of predicted earnings (based on pre-MBA characteristics and MBA performance). But they do not investigate this point further and do not assess the role of selection as a determinant of the evolution of the gender wage gap.

The issue of labor market selection is obviously a classic one within labor economics, dating back at least to Heckman (1974). But there is only a relatively small literature on the effect of differential selection across genders on the gender wage gap, and this literature focuses exclusively on the secular evolution of the gender wage gap. Blau and Kahn (2006) use the PSID to analyze the determinants of the slowdown in the reduction of the gender wage gap between the 1980s and 1990s, and argue that a reduced rate of growth in the positive selection of women in the 1990s is partly behind the phenomenon. Mulligan and Rubinstein (2008) use repeated CPS cross sections and a Heckman two-

[1] Goldin et al. (2017) seek to explain how much of the growth in the wage differential between working women and men can be explained by differential dynamic sorting across establishments with different mean earnings. Its contribution is in the tradition of Abowd et al. (1999), who decompose wages into worker and firm components and analyze the sorting of good workers into good firms, and Card et al. (2016), who analyze the importance of this type of sorting to explain gender wage gaps in Portugal. The focus of Goldin et al. (2017) is complementary to ours, since we analyze selection into work participation and the overall composition of the pool of working women and men, irrespectively of their sorting across firms. They explain part of the evolution of the wage gap among women and men continuously employed through their sequential employment in firms with different productivity levels. We explain part of the age profile of the gender wage gap by pointing out that *different pools* of women and men – in terms of the underlying distribution of ability – are employed at each age.

step estimator to show that selection of women into the labor market changed from negative in the 1970s to positive in the 1990s. They argue that improved female selection into the labor force is one of the main reasons behind the observed reductions in the conditional gender wage gap during this period. Herrmann and Machado (2012) perform regressions of participation on measures of cognitive ability (test scores) separately for men and women and look at the evolution of differential selection on ability across genders over time.

The main difference between our paper and the literature discussed in the previous paragraphs is that we focus on the role of selection as a determinant of the evolution of the gender wage gap over the lifecycle.[2] In addition, most of the research on the evolution of the gender wage gap within birth (or graduation) cohorts focuses on highly-educated workers in very competitive careers in developed countries (e.g., Corcorant et al., 1993; Goldin and Katz, 2008; Bertrand et al., 2010), while we look at workers of all skill levels in all occupations in the context of a developing country.

The remainder of the paper is organized as follows. Section 2 presents a simple theoretical model that motivates our empirical exercise and helps to guide our interpretation of the results. Section 3 describes the data. Section 4 performs our descriptive exercise accounting for the role of selection in the early career growth of the gender wage gap. Section 5 presents the methodology and the results related to our exercise on the lifecycle relationship between ability and formal employment. Section 6 offers concluding remarks.

2. Theoretical Background

The most common measure of labor market discrimination, plagued by problems of unobservable confounding factors, comes from the estimation of a wage equation such as

$$ln\, w = \alpha + \rho s + \gamma_0 x + \gamma_1 x^2 - \theta f + \boldsymbol{z}'\beta + \varepsilon, \tag{1}$$

where w denotes hourly wages, s level of schooling, x labor market experience, f a dummy variable indicating gender (female), $\boldsymbol{z}$ a vector of demographic variables correlated with wages, and ε a random term. In this setting, and abstracting from the various potential limitations from this approach, θ is interpreted as the wage differential between men and women for given observable characteristics. It

[2] Adda et al. (2017) look at the timing of fertility and how it is related to changes in the ability composition of working women, but do not analyze the implications of this change to the evolution of the gender wage gap. Our paper can be seen as complementary to their work: while they look at the determinants of the change in the ability composition of women in the labor market, we look at the implications of this compositional change for the evolution of the gender wage gap. Machado (2013) proposes and implements an IV-inspired estimator for the gender wage gap that is robust to arbitrary selection into the labor market. Her work is related to the point discussed here, but does not focus on the nature of selection or on its evolution over the lifecycle.

is therefore commonly taken as an indicator of the degree of labor market discrimination across genders.

The most widely used justification for this empirical specification comes from a lifecycle model of earnings (Mincer, 1974) extended to incorporate employer discrimination as defined by Becker (1957). In order to highlight the limitations of this framework and to show how its shortcomings motivate and guide our empirical exercise, we quickly summarize this extended Mincer model here (following the presentation from Cahuc and Zylberberg, 2004). Consider an individual born in period 0 who attends school until age *s*, when she enters the labor market. Her working life ends at period *T*, when she retires. During the "schooling period" [0, *s*], time is allocated exclusively to human capital accumulation. During working life (*s*, *T*], the individual decides on how much time to devote to training (which further increases human capital) and to work.

Let $p(t) \in [0,1]$ be the fraction of time allocated to training at instant $s + t$, with $[1 - p(t)]$ indicating the fraction of time allocated to work, where $0 < t \leq T - s$. Training increases the worker's stock of human capital, $h(\cdot)$, according to differential equation $\dot{h}(s+t) = \rho p(t) h(s+t)$, $\forall\ t \in [0,T - s]$, where ρ is the rate of return to training after leaving school. Within the context of the Mincer model, the positive effect of labor market experience on wages is interpreted as being associated with the accumulation of human capital through the time investment *p(t)*. In other words, time working in the market in the past is associated with a higher wage in the future because past work is related to direct accumulation of human capital through on-the-job training.

With perfect competition in the labor market, wages at instant $s + t$ are given by

$$y(s+t) = A[1 - p(t)]h(s+t), \tag{2}$$

where A is a productivity constant. From equation (2), the individual's earnings are proportional to her stock of human capital, which reflects both schooling and labor market experience, and to the labor supplied in the current period $(1 - p(t))$. In this context, if all employers discriminate against women in the labor market, with a common discrimination coefficient θ (see Becker, 1957), equilibrium wages will be such that $y_f(s+t) = (1-\theta)y_m(s+t)$ for equally productive men and women, where subscripts denote genders.

Integrating the human capital accumulation equation from $t = 0$ to $t = x$ leads to the expression $h(s+x) = h(s)e^{\rho \int_0^x p(t)dt}$. Substituting into equation (1), this yields $y(s+x) = (1 - I(f)\theta)A[1 - p(x)]h(s)e^{\rho \int_0^x p(t)dt}$, where *I(f)* is an indicator function equal to 1 if the individual is female. That is, the income of an individual with *x* years of experience depends on her stock of human capital upon leaving school ($h(s)$) and on the additional human capital accumulated in the marketplace ($\int_0^x p(t)dt$). Mincer (1974) makes the simplifying assumption that the fraction of time spent on training declines linearly with x: $p(x) = p_0[1 - (x/T)]$. Using this, substituting $h(s) = h(0)e^{s\rho}$, where $h(0)$ represents the innate stock of human capital, and taking natural logarithms on both sides of equation (2) leads to:

$$ln\, y(s+x) = ln\, Ah(0) + \rho s + \rho p_0 x - \rho\left(\frac{p_0}{2T}\right)x^2 + ln[1 - p(x)] + ln(1 - I(f)\theta). \tag{3}$$

Using the fact that $ln(1 - I(f)\theta) \sim - I(f)\theta$ and assuming that $Ah(0)$ is a linear function of demographic characteristics $\boldsymbol{z}$ and a random term ε, such as in $ln\, Ah(0) = \alpha + \boldsymbol{z}'\beta + \varepsilon$, one can rewrite this expression as a traditional Mincer equation:

$$ln\, w(a) = \alpha + \rho s + \gamma_0 x + \gamma_1 x^2 - \theta f + \boldsymbol{z}'\beta + \varepsilon,$$

where $w(a) = y(s + x)/[1 - p(x)]$ is the hourly wage at age a (with $a = s + x$), $\gamma_0 = \rho p_0$, and $\gamma_1 = -\rho(p_0/2T)$, and f is the dummy variable for gender defined before.

If, in addition, gender discrimination varies with age and, conditional on schooling and other characteristics, there is a separate competitive labor market for individuals of different ages, this equation would be rewritten as:

$$ln\, w(a) = \alpha + \rho s + \gamma_0 x + \gamma_1 x^2 - \theta_a f_a + \boldsymbol{z}'\beta + \varepsilon, \tag{4}$$

with θ_a indicating the age-specific coefficient of discrimination against women, and f_a a dummy variable equal to one for women of age a.[3] This would be an adequate specification if increases in wages with experience (x) reflect career advancements associated with promotions to higher level occupations, for which it might be reasonable to assume that gender discrimination would be higher.

This model provides the simplest theoretical motivation for estimating a Mincer equation where the log of hourly wages is regressed on a linear term on schooling, a quadratic function of experience, age-specific gender dummies, and a set of demographic controls. This formulation makes it clear that the demographic controls $\boldsymbol{z}$ are intended to represent determinants of individual productivity in investments in human capital, such as its initial stock (summarized by $h(0)$, corresponding to ability and family background) and other factors affecting the return to productive attributes in the labor market (summarized by A).

The Mincer model does not allow for labor supply decisions at the extensive margin. This is relevant in our context because differential labor market entry and exit across genders present two potentially serious challenges for the interpretation of θ_a as capturing the evolution of gender

[3] In principle, the same argument could be made for estimating the gender wage gap by educational level, for example. As the focus of the paper is the evolution of the gender wage gap over the professional lifecycle, we only model explicitly the age heterogeneity.

discrimination. First, actual experience x is rarely observed in the data, so empirical specifications traditionally use potential experience, defined as $x_p = age - s - 6$. As recognized by Mincer and Polachek (1974), x_p is a particularly poor predictor of actual experience for women given their more frequent interruptions in labor market activity. This artificially inflates the coefficient on the gender dummies.

Second, measures of ability are also rarely observed in the data. If controls for ability related to $h(0)$ are not included in $\boldsymbol{z}$, heterogeneous participation across genders may bias the coefficient on the gender dummy, since the pool of women can be potentially different from the pool of men in terms of unobserved skills. Moreover, the strength and direction of this differential selection may vary across genders over the lifecycle, in which case it will impact the estimated age profile of the gender wage gap. If selection becomes stronger for men than for women over time, the mean ability of men will increase in relation to that of women. Without accounting for selection, this differential trend will be reflected on an increasing gender wage gap as individuals age. On the other hand, if female selection becomes increasingly more positive with age as compared to that of men, the true growth in the gender wage gap will be even larger than that usually estimated from standard Mincer regressions.

The first of these problems is the object of the large literature discussed in the introduction, which includes Oaxaca and Regan (2009), Bertrand et al. (2010), and Blau and Kahn (2013). We focus here on the second problem, which has been studied from the perspective of secular changes in the gender gap but not of its lifecycle evolution (Blau and Kahn, 2006; Mulligan and Rubinstein, 2008; Herrmann and Machado, 2012; Machado, 2017).

If one has access to repeated observations on the same workers, individual fixed effects can be included as part of the demographic controls $\boldsymbol{z}$ to overcome the problem of differential selection on ability. Since $h(0)$ does not change with time, it seems reasonable to assume that it would be captured by the individual fixed effects, which could then be interpreted as proxies for the set of time-invariant skills valued by the labor market. If, additionally, it is possible to build an adequate measure of actual labor market experience, one could take care of both problems alluded to above. Under these conditions, OLS estimation of equation (4) would deliver consistent estimates of the parameters of interest, including the evolution of the gender wage gap and the measure of ability represented by the individual fixed effects. The latter could then be used to explicitly analyze the pattern of labor market selection over the lifecycle.

This is the strategy we adopt in this paper. Put simply, we ask how allowing for a time invariant additive measure of individual ability in a specification similar to equation (4) changes the conclusions related to the evolution of the gender wage gap over the lifecycle (θ_a). Following, we ask what this measure of individual ability tells us about the change in the relationship between ability and employment across genders over time. More general formulations of the human capital production function could lead to specifications where ability would affect, for example, the returns to schooling or experience in the wage equation. Broader interpretations of discrimination would account for the anticipated effects of discrimination on human capital investment decisions, therefore incorporating the effect of θ_a on decisions over s as part of the total effect of discrimination on the gender wage gap. Our objective here is less ambitious than that. We simply ask how allowing for an additive individual fixed effect in a wage equation such as (4) changes the conclusions related to the growth of the gender wage gap through the lifecycle and, more generally, our understanding of the dynamic pattern of selection into the labor market across genders. This straightforward formulation is theoretically consistent with the Mincer model presented in this section and seems like a natural first step in this direction.

3. Data

3.1. Sample and Construction of Variables

We use the RAIS (*Relação Anual de Informações Sociais*) dataset, a very large restricted-access administrative record collected by the Brazilian Ministry of Labor. RAIS is a longitudinal employee dataset covering the universe of formal employees in Brazil. Every year, tax-registered firms are legally required to report every worker formally employed at some point during the previous calendar year. Each worker in the dataset is identified by a unique national social insurance number (PIS, *Programa de Integração Social*), which is similar to a social security number. This allows us to follow workers over time and across firms.

For each year, the dataset includes: (i) firm-related variables, such as sector of activity, size, state and municipality; (ii) worker-related variables, including gender, age and schooling; and (iii) job-related variables, such as monthly earnings, occupation, contracted hours of work (weekly), 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 a given year, information about the month of hiring is also provided.

RAIS is very large, with more than 55 million observations only in 2010. Since working with the full dataset is not feasible computationally, we use a random sample of workers. First, we collect the identification numbers of all workers born in 1974 who appear in the RAIS dataset at some point between 1995 and 2010. Then, we use a random sample of 30% of these identification numbers and search for each of these workers in all years from 1995 to 2010. The resulting dataset contains the complete 1995-2010 formal work history of each individual in the sample.

By construction, all workers are in the 21-36 years old range. The advantage of using only one birth cohort is that our subsequent analysis is not confounded by cohort effects. In addition, individuals born in 1974 were relatively young (21 years old) in 1995, the first year available in our dataset. So their (unobserved) work history prior to 1995 is unlikely to be either long or important enough from the perspective of human capital accumulation.

We also apply some filters to the data. We drop all workers without valid identification numbers (PIS) and with negative earnings. We discard all observations with weekly working hours lower than 5 or higher than 60. We only keep the 'main job' held by each individual in a given year, defined here as the job with the highest average real monthly earnings. Finally, in order to be able to estimate individual fixed effects in the regression analyses, we discard individuals who appear only once between 1995 and 2010 (we also deal explicitly with issues raised by the precision of fixed effects estimates in our empirical strategy).[4] The resulting dataset is composed of 443,392 workers, of which 44.1% are women. The total number of observations (workers × years) is 3,639,146.

Our wage variable is *lwage*, the logarithm of average real monthly earnings.[5] We classify workers into four schooling groups, according to their final educational attainment (highest schooling level observed in the data): less than high school, complete high school, incomplete college, and complete college (since we use individual fixed effects in our main specifications and we need changes in educational levels to identify the coefficients on schooling, there is no gain in including dummies for lower levels of schooling). We also generate sets of dummy variables for age (***age***), sector of activity of the firm (***sector*,** a vector of dummy variables for 26 aggregate sectors of economic activity), firm size (***size***, a vector of dummies for 10 categories of firm size, as measured by the number of employees), and state (***state***, a vector of dummies for the 26 Brazilian states, plus the Federal District).

We construct several variables measuring labor market experience that are included in a vector ***exper***, which describes previous formal labor market experience. Our goal is to be as flexible as possible in characterizing the returns to formal labor market experience. Following Spivey (2005), we build four non-parametric experience vectors: ***empl, FTempl***, ***MYempl***, and ***FTMYempl***. Each of these

[4] As Fernandes (2013), we adopt a set of algorithms to fix some data inconsistencies in variables such as schooling and age. Details about these inconsistencies and our algorithm are described in Appendix A1.

[5] The average of earnings is taken over all months of the year in which the contract was active. Real earnings are in *reais* of December 2010 (average exchange rate of 1.69 R$/US$), calculated by deflating nominal earnings of each month using the Brazilian consumer price index IPCA (*Índice de Preços ao Consumidor Amplo*).

vectors has dummies referring to previous labor market experience, up to a total of 15 years (for the last year observed). For example, ***empl*** in year *t* contains *t – 1995* dummies (*empl(1)*, *empl(2)*, …, *empl(t-1995)*), where *empl(k)* indicates whether the individual held a formal job for any length of time *k* years in the past. The other three vectors are analogous to ***empl*** but refer to stronger types of labor market attachment. We define working "most-year" as working for at least 9 months in a given year and "full-time" as at least 35 hours per week. We use combinations of these definitions of full-time and most-year employment in the formal labor market to create the other dummy variables indicating the status of workers in a given year: full-time (FT), most-year (MY), and full-time most-year (FTMY).[6]

Finally, the variable used in the second step of our empirical analysis is a dummy for full-time most-year employment. There are four such variables, referring to ages 21, 26, 31, and 36, trying to capture the evolution of employment during the first 15 years of professional life.

In some robustness exercises, we also use data from the Brazilian National Household Survey (PNAD, *Pesquisa Nacional por Amostra de Domicílios*), conducted by the Brazilian Census Bureau (IBGE, from *Instituto Brasileiro de Geografia e Estatística*). The PNAD is a nationally representative survey covering all Brazilian states and providing demographic information typically available from labor market surveys, including employment by labor market status (formal and informal employees, and self-employed, in addition to employers).

3.2. Descriptive Statistics

The simple conditional gender wage gap estimated with the usual controls in our sample starts at 0.15 log point at age 21, rising steadily by 0.19 over the following 15 years to reach 0.34 by age 36. This profile documents the well-known pattern of early career growth in the gender wage gap in our data. It echoes previous findings in the literature both in Brazil and elsewhere, such as illustrated by Bertrand et al. (2010), Li and Miller (2012), and Fernandes (2013).

In order to assess the representativeness of our sample, which is based on administrative records and comprises only formal sector workers, we also calculate analogous gender wage gap profiles with data from the Brazilian National Household Survey (PNAD). We try to replicate the same exercise conducted with RAIS by following a single cohort over time. Since the PNAD has a much smaller sample, we look at individuals born between 1973 and 1975, and follow these cohorts in the years corresponding to our RAIS sample (from 1995 to 2010, with the exceptions of 2000 and 2010, when the PNAD survey was not conducted). We estimate the exact same specification of the Mincer

[6] We use nine months of employment to characterize full-year employment due to very high turnover observed in the Brazilian labor market (Corseuil et al., 2013).

regression used with the RAIS dataset to recover the conditional gender wage gaps, restricting the sample to full-time workers and regressing the logarithm of real monthly wages on educational level (dummies), age (dummies), hours of work, tenure, sector of economic activity (dummies), and state of residence (dummies). We restrict this specification to be as comparable as possible across the two datasets (this is not the specification we use later on when conducting our main analysis with the RAIS dataset).

When we look at all employees – including formal and informal – in the PNAD data, we estimate a conditional wage gap that starts at 0.18 log point at age 21 and rises by 0.16 to reach 0.34 by age 36. If we restrict the PNAD data to formal workers, we estimate a wage gap of 0.17 at age 21 and 0.31 at age 36. As mentioned before, the gender wage gap estimated from the RAIS dataset under this specification starts at 0.15 at age 21 and grows to 0.34 by age 36. The average gender wage gap between ages 21 and 36 is 0.25 log point in both PNAD samples, and 0.26 in the RAIS sample. Overall, it seems fair to say that the data from RAIS portrays very closely the gender wage gap among all employees (formal and informal) in the Brazilian labor market.

Figure 1 plots the formal employment rate by age for each gender from the RAIS dataset. Formal employment rises continuously for both men and women between ages 21 and 36, but the levels are always higher for men. The difference in employment rates starts at 5 percentage points, reaches almost 8 percentage points by the mid-20s, and then falls back to roughly 3 percentage points by age 36. The profile of rising formal employment during the early stages of professional life in Brazil has been documented before in the literature (see, for example Cruces et al., 2012). But it is important to understand exactly what this figure means. There is a very high degree of employee turnover in Brazil, even in the formal sector. So rather than indicating that more individuals over time are entering the formal sector and that those who enter tend to remain there indefinitely, Figure 1 shows that the difference between the rate of formal labor market entry and exit increases more for men than for women over this age interval. For both genders, it remains true that exit and turnover rates are very high throughout.

To illustrate this point, Figure 2 restricts the sample to individuals who held a formal job at ages 21, 26, and 31, and looks at their survival rate into formal employment up to age 36. By construction, this normalizes employment rates to 100 percent for both genders for these three ages. The figure shows that, irrespectively of the starting point, there are substantial reductions in formal employment rates, meaning that a sizable number of men and women leave the formal labor market at some point as time goes by (the figure includes reentries after initial exits). For those who held a formal job at age

21, survival rates by age 30 are only 40 percent for females and 50 percent for males. Starting at later ages, the initial drop in employment is not so steep and yearly survival is somewhat higher, but the same qualitative pattern holds. For individuals who held a formal job at age 31, for example, only 62 percent of men and 55 percent of women still hold a formal job at age 36. This large turnover, with sizeable differences across genders, opens up the possibility of substantial changes in the composition of the pool of working men and women and leaves room for selection to affect the evolution of the gender wage gap. Though men's survival rate falls initially almost as quickly as women's, it stabilizes at a much higher level, close to 10 p.p. above that of women when we look at individuals employed at age 21.[7]

Table 1 presents descriptive statistics for some of the variables used in the empirical analysis. Average earnings for men (R$ 1,455 in Brazilian *Reais* of December 2010, which corresponded to US$ 861) are 18.7 percent higher than for women (R$ 1,226, or US$ 725). The variance of earnings is also larger for men. Men account for 59 percent of the worker × year observations. Average age is slightly higher for females (29.3 *vs.* 28.9), while average working hours are higher for males (42.6 *vs.* 40.5), consistent with the documented pattern that women start working later and typically work fewer hours. The table also shows that women in the sample are relatively more educated than men, with a lower share with less than high school education (30.7 percent vs. 47.4 percent) and a higher share with college (20.4 percent *vs.* 11.2 percent).

We have a representative sample of the universe of individuals born in 1974 that held a formal job at some point between 1995 and 2010. Our sample is not representative of the entire 1974 cohort since formal jobs are not randomly distributed within this cohort and not all individuals born in 1974 held a formal job at some point during the 16-year period we consider. Nevertheless, since we consider individuals who were formally employed at any point between 1995 and 2010, our sample is closer to the overall profile of this cohort than to the profile of individuals formally employed at a given moment in time. This should be expected, given the high degree of employment turnover mentioned before.

We illustrate this point with the last two panels in Table 1, which present the distributions of levels of schooling in the 2010 census for men and women born in 1974. We also present the same

[7] In addition, men are more likely than women to transition from formal employment to informal or self-employment, so the difference in survival rate in overall employment is larger than that portrayed in the figure. According to the Brazilian Monthly Employment Survey (*Pesquisa Mensal de Emprego*), for example, women leaving formal employment between ages 21 and 36 have a probability of roughly 18 percent of working in the informal sector or self-employed in the following year, while the analogous number for men is around 30 percent (authors' calculations based on data for the month of March from the 2002-2010 Monthly Employment Survey).

distribution for individuals from this cohort who were formally employed in 2010. Among individuals born in 1974 who were formally employed in the 2010 census, 63 percent are men and 37 percent are women, while our sample contains 56 percent of men and 44 percent of women. In the overall population from this cohort in the 2010 census, we obviously have 50 percent of men and 50 percent of women. In terms of the educational distribution, our sample, not surprisingly, over represents higher levels of schooling when compared to the overall population, and particularly so for women.

4. Accounting for Selection in the Early Career Growth of the Gender Wage Gap

In this section, we adopt a descriptive strategy to assess the contribution of differential selection across genders to the early career growth in the gender wage gap in the 1974 cohort. We do this by estimating the lifecycle profile of growth in the gender wage gap using two different sets of wage regressions: a simple OLS specification akin to the one commonly used in the literature and an alternative specification that controls for individual fixed effects.

Our focus is on the age profile of the gender wage gap. So, we include in the wage regressions a dummy for males and interactions of this male dummy with age dummies. For the regression including individual fixed effects, we cannot identify the male dummy separately from the individual fixed effects. But we can still include interactions of age dummies and the male dummy to recover the profile of growth in the gender wage gap over the lifecycle. A comparison of this profile across the OLS and the fixed effects specifications tells us how much of the initial growth in the gender wage gap is due to differential changes in selection across genders. This is the comparison we undertake.

The regressions have as dependent variable the logarithm of real monthly earnings, *lwage*. We control for traditional individual and job characteristics, such as schooling (non-parametrically in the four levels of education listed in Table 1: less than high school, complete high school, incomplete college, and complete college), weekly contractual hours of work, and state and age dummies (since we are looking at a single cohort and have age dummies, there is no need to control for time as well).[8] To account for firm characteristics correlated with wages, we control for sector of activity (dummies) and firm size (dummies for categories of number of workers), which can be seen as proxies for firm

[8] We use the logarithm of monthly wages as dependent variable and control for hours, rather than using hourly wages as dependent variable, because we only observe contractual hours in the RAIS dataset. In this situation, including contractual hours as a control is typically thought to be the best option. In addition, labor supply in the intensive margin may also affect wages, in which case this specification should also be preferred (Blau and Kahn, 2000). We restrict the educational dummies to the four categories mentioned in the text because we use individual fixed effects and thinner educational categories in the RAIS data are measured with a lot of error. When we include thinner educational categories in our specifications with individual fixed effects, estimates become very imprecise.

productivity.[9] Importantly, we also control for tenure in the current job (number of months in present job, as of December of year t or up to the separation month) and for the broad array of variables measuring labor market experience explained in Section 3.1, which describe non-parametrically and in detail the previous employment history of individuals (vector ***exper***).[10] The inclusion of tenure on the current job as an additional independent variable controls, to some extent, for previous labor market experience before age 21, as long as this experience comes from continuous employment in the same firm (notice that formalization rates before age 21 are very small, so in any case this should not be a first order concern).

Our specification for the wage equation does a better job than most of the literature in controlling for past experience. As discussed in Section 2, the literature traditionally uses potential experience as a proxy for actual experience, which confounds actual experience with past non-working periods. Not taking into account a full measure of past experience tends to produce biased estimates of all coefficients in wage regressions (Blau and Kahn, 2013). By contrast, we fully characterize each individual's work history as recorded in the RAIS dataset. Since we are looking at individuals who are aged 21 in the first year of the dataset, we have a close to complete description of their formal labor market histories.

The results are reported in Table 2. For purposes of comparison, Column 2 presents the conditional gender wage gap by age estimated from a simple OLS regression without experience controls. It reproduces patterns documented before in the literature, both in Brazil and elsewhere (see, for example, Blau and Kahn, 2000; Ñopo, 2012; Fernandes, 2013). The conditional gender wage gap for the 1974 cohort started at 13 percent at age 21, reached 27 percent at age 29, and peaked at 33 percent at age 36. The corresponding cumulative growth in gender wage gap from age 21 onwards is listed in column 3, indicating an increase of 20 percentage points (more than 150% of the initial level) during the first 15 years of this cohort's professional life. Just in the first eight years between ages 21 and 29, the conditional gender wage gap increased by 100%. This relatively short period during the early career concentrated most of the lifetime growth in the gender wage gap.

[9] Ideally, we would want to control for firm fixed effects as well in this specification. But we only have a relatively small sample of the overall population employed in the Brazilian formal labor market. So the overlap of firms and individuals is quite limited. If we were to include firm fixed effects, we would need to give up our focus on a single cohort over time. Instead, we choose to control for firm size and sector of economic activity.

[10] Notice that, since we are following a single cohort over time, we cannot distinguish between period and age effects. Though the profile of the gender wage gap that we recover from the data is similar to that documented in the literature in other settings, we claim only to be describing the experience of one specific cohort.

Columns 4 and 5 display numbers analogous to those from columns 2 and 3, but for specifications that control for previous labor market history. Surprisingly, there is very little change in the growth profile of the gender wage gap as we compare columns 4 and 2. There is a mild reduction of the gender wage gap as we control for labor market history, of the order of 6 percent on average. This is in strong contrast to evidence from the US high-skilled labor market, where actual experience has been shown to account for a major part of the gender wage gap (Bertrand et al., 2010). Controlling for previous experience also leads only to a mild reduction in the profile of growth of the gender wage gap. Column 6 calculates the difference between columns 3 and 5, and column 7 presents this same difference as a percentage of the value from column 3. Previous experience explains part of the growth in the gender wage gap, particularly so during the first ten years of professional life, but only 6.6 percent by age 36.

Interpreting these results as indicating that our experience variables do not capture real labor market experience would be a mistake. The joint F-statistic for our set of experience variables (listed in the table) is above 1,000. The experience variables are strongly significant and indicate substantial returns to previous employment, even though they cannot account for a major part of the gender wage gap or of its growth over time. Figure 3 illustrates the relevance of our experience variables for earnings by portraying the returns to continuous formal employment (and full-time most-year employment) by age. The return to one year of previous employment in the 1974 cohort was, on average, 2.6 percent. Being employed full-time most year for one additional past year was associated with wages 3 percent higher, on average. By age 26, continuous employment without most-year and full-time attachment over the previous 5 years was associated with wages 13 percent higher, while continuous most-year full-time employment was associated with increases of 29 percent in wages. Individuals at age 36 experienced wages close to 44 percent higher from continuous full-time most-year employment in the previous 15 years. Our experience variables thus seem to do a good job in capturing returns to actual formal labor market experience.

Column 8 in Table 2 presents the cumulative growth in the conditional gender wage gap estimated from the fixed effects specification, where we control for individual unobserved factors. The yearly growth in the gender wage gap estimated with fixed effects is substantially smaller than that from column 5. The difference between columns 5 and 8 can be directly interpreted as the share of the early career growth in the gender wage gap that can be attributed to differential selection, conditional on previous labor market experience. Figure 4 illustrates this point graphically by plotting columns 3, 5 and 8. The difference between the thicker (green) and the dashed (blue) curves in the figure represents the role of formal labor market experience in the early growth of the gender wage gap, while the difference between the dashed (blue) and the thinner (red) curves represents the role of changing selection over the lifecycle.

Column 9 in Table 2 calculates the difference between columns 5 and 8, and column 10 presents this same difference as a percentage of the values from column 5. Changes in selection across genders explain an important share of the growth in the gender wage gap, but the role of selection is not monotonic across ages. Selection played a more important role in the first years of the professional career, before the 30s, when it accounted for a major part of the growth in the gender wage gap (above 50 percent). As time went by and individuals aged, this role became relatively less relevant. By age 36, differential selection accounts for 32 percent of the cumulative growth in the gender wage gap.

A nontrivial portion of the early career growth in the gender wage gap for the 1974 cohort – after accounting for education, experience, and other observables – was due to the compositional change in the pool of women and men being compared to each other at each point in time. The fact that this mechanism seems to be more important at earlier ages supports the idea that it may be partly driven by changes in marital status and fertility decisions. In the case of the 1974 cohort in the Brazilian formal labor market, selection turns out to be a much more important factor than past labor market experience, which has been the focus of most of the recent international literature.

5. Ability and Employment over the Early Professional Life

5.1. Empirical Strategy

This section proposes a methodology that takes advantage of the panel structure of the RAIS dataset to describe how the relationship between ability and formal employment evolves during the early years of professional life. The goal is to shed light on the compositional changes behind the results obtained in the previous section. In order to achieve this goal, we proceed in two steps. First, we build measures of unobserved ability based on individual fixed effects estimated from wage equations. Following, we use these measures of ability to investigate the relationship between ability and employment during the early career.

5.1.1. First Step: Worker Fixed Effects as a Measure of Ability

The crucial step in our empirical strategy is to recover a measure of ability to be used in the estimation of employment regressions. Our longitudinal dataset allows us to estimate wage equations separately for each gender, including individual fixed effects (FEs) and controlling for other observable characteristics. We interpret the estimates of worker FEs as pecuniary measures of the set of time-invariant unobserved abilities that are valued by employers, such as cognitive skills, commitment, motivation and other "soft skills."

We estimate two equations separately by gender to be as flexible as possible and to allow the returns to productive attributes to vary between men and women. This eliminates the possibility that individual fixed effects partly capture gender discrimination, since we let the coefficients on observable characteristics and the average of the distribution of fixed effects to be gender-specific. But, at the same time, it introduces non-trivial issues in the comparison of fixed effects across the two groups. This point is discussed in detail later in this section.

We start by estimating wage equations separately for each gender:

$$lwage_{git} = \gamma_{gi} + \boldsymbol{\delta_{1g} exper_{git}} + \boldsymbol{\delta_{2g} educ_{git}} + \beta_{1g} tenure_{git} + \beta_{2g} hours_{git} + \boldsymbol{\alpha_{1g} sector_{git}} + \boldsymbol{\alpha_{2g} size_{git}} + \boldsymbol{\alpha_{3g} state_{git}} + \boldsymbol{\alpha_{4g} age_{t}} + \varepsilon_{git}, \qquad (5)$$

where g indexes gender (f or m), i indexes worker, and t indexes year. The estimated fixed effects γ_{gi} are used to construct the measure of ability used in the second step of our analysis.

In this context, it is essential to adequately control for professional history. Otherwise, the individual fixed effects will capture not only individual ability but also accumulated labor market experience, biasing the results. We control nonlinearly for experience using the ***exper*** vector described in Section 3.1, which details the previous formal employment history of individuals. In a robustness exercise, we also incorporate potential informal labor market experience – from self-employment and unregistered employment – in this vector. Additional controls include the vector of dummies for educational levels (***educ***), tenure in the current job (*tenure*), contracted hours of work (*hours*), and dummies for sector of activity (***sector***), firm size (***size***), state (***state***), and age (***age***). The discussion in Section 4 about the choice of controls applies to this specification as well. Since we are looking at a single cohort and have age dummies, there is no need to control for calendar year. Finally, ε is an error term assumed to be orthogonal to the explanatory variables.

We deal explicitly with the main potential sources of bias in the interpretation of individual fixed effects as measures of unobserved ability. First, as mentioned before, the RAIS dataset only allows us to construct measures of experience in the formal labor market, leaving out unregistered employment and self-employment spells, which may be relevant in the case of Brazil. We address this concern in our robustness exercises by including measures of potential unregistered employment and self-employment experience as additional controls when estimating equation (5). These measures of potential informal experience are estimated with data from the Brazilian National Household Survey

(PNAD) by age-gender-state-education cell and are discussed in further detail when they are introduced in the results section.

Another potential concern is that entry and exit driven by other factors may confound temporary labor market shocks with unobserved individual ability. For example, if there is a group of women who enter the labor market only in periods of particularly heated economic activity and exit afterwards, the model could in principle assign to unobserved individual ability an effect that is in reality driven by a local economic shock (this could happen depending on the geographic nature of the economic shock and on the set of controls included in the regression, and could lead to spurious correlation between measured ability and employment). We address this possibility in our robustness exercises by re-estimating equation (5) using a Heckman two-step procedure to correct for time-varying selection into the labor market. In this specification, we use the formal employment rate by gender-age-state-education cell as the determinant of formal employment excluded from the wage equation. We discuss the implementation of this exercise in detail when presenting the results.

Normalization and Interpretation of Worker Fixed Effects

The fixed effects γ_{gi} in equation (5) are worker-specific time-invariant factors that are added to workers' earnings in every period when they are formally employed. We interpret γ_{gi} therefore as the monetary value of the set of skills that is valued in the labor market.

The dependent variable *lwage* in equation (5) is the logarithm of earnings, so γ_{gi} gives the approximate percentage increase in earnings due to worker ability. For example, consider two workers, A and B, with $\gamma_{gA} = 0.3$ and $\gamma_{gB} = 0.1$. If A and B had the exact same set of observable characteristics (such as schooling, experience, state of residence, etc.), A would command earnings approximately 20% higher (in expectation) than B solely because of higher ability. However, we fit our model separately by gender, which makes comparisons of γ_{gi}'s valid only within genders. It is correct to say that a woman A with $\gamma_{fA} = 0.5$ is more skilled than a woman B with $\gamma_{fB} = 0.3$, but we cannot say that she is more skilled than a man C with $\gamma_{mC} = 0.3$, since the estimates of γ_{gi} across genders come from different distributions.

In our second step, we are interested in comparing worker fixed effects across genders. In order to allow for this comparison, we normalize fixed effects within each gender. The normalization re-centers and rescales the fixed effects distribution for each gender so as to make the two distributions, in principle, comparable. Let g be a specific population group. Normalized fixed effects are given by $\hat{\gamma}_i^s = \frac{\hat{\gamma}_{gi} - m_g}{\sigma_g}$, where m_g and σ_g are, respectively, the mean and standard deviation of estimated fixed effects for individuals in group g. The normalized estimates $\hat{\gamma}_i^s$ are then used instead of the original $\hat{\gamma}_{gi}$ in the second step of our empirical strategy.

This normalization obviously changes the interpretation of the fixed effects. Fixed effects now measure the distance (in standard deviations) between the individual's ability and the mean of the relevant fixed effects distribution. They no longer have a direct monetary interpretation.

Under the assumption that the underlying distribution of pre-market ability is the same across the universe of men and women who appear in the RAIS dataset at some point between 1995 and 2010, this procedure allows relative comparisons across genders to be made. In other words, this assumption implies that – after controlling for personal characteristics and allowing these characteristics to be priced differently across genders in the market, and purging the differences in the level and dispersion of pay across men and women – the underlying distribution of ability for individuals who held a formal job at some point between 1995 and 2010 would be the same across genders. In this context, inter-gender comparisons of the effect of ability on employment have some meaning. For instance, if man A and woman B have $\hat{\gamma}_A^s = \hat{\gamma}_B^s$, then their positions on their respective ability distributions are the same. If, in addition, A has a probability of employment higher than B at a certain age, we can say that the probability of employment for that (normalized) ability is higher for men than for women in that age group.

In any case, even if the underlying distributions of abilities are not the same across genders, this comparison is still informative of the relative compositional changes taking place across formally employed men and women as individuals age.

5.1.2. Second Step: Ability and Employment

In our second step, we estimate regressions for full-time most-year employment in which the main explanatory variable is the normalized individual fixed effect, $\hat{\gamma}^s$, and its interaction with a gender dummy, $male \times \hat{\gamma}^s$. The goal is to understand the pattern of selection into employment over the lifecycle and how it varies across genders. The coefficient on $\hat{\gamma}^s$ in these regressions is interpreted as measuring the sign and strength of selection into employment, while the coefficient on $male \times \hat{\gamma}^s$ measures the gender differential in selection. We focus on how selection into employment varies across ages by running separate regressions for our measure of full-time most-year employment at ages 21, 26, 31, and 36. The main specification is the following:

$$empl_{ai} = \theta_{0a} + \theta_{1a}\hat{\gamma}_i^s + \theta_{2a}male_i \times \hat{\gamma}_i^s + \boldsymbol{\theta_{3a}educ_{ai}} + \boldsymbol{\theta_{4a}state_{ai}} + \theta_{5a}male_i + \boldsymbol{\theta_{6a}}male_i \times \boldsymbol{educ_{ai}} + \boldsymbol{\theta_{7a}}male_i \times \boldsymbol{state_{ai}} + \mu_{ai}, \quad (6)$$

where a indexes age ($a \in \{21, 26,31,36\}$) and i indexes worker. This same specification is estimated separately for each age a. The employment variable *empl$_a$* indicates whether the individual worked full-time most-year at age a. Variable $\hat{\gamma}^s$ corresponds to the normalized fixed effects obtained from the estimation of the Mincer regression in the first step, and $male \times \hat{\gamma}^s$ is its interaction with a dummy for males. We control for schooling dummies ($\boldsymbol{educ_a}$) and state dummies ($\boldsymbol{state_a}$), and also include interactions of these variables with the *male* dummy, $male \times \boldsymbol{educ_a}$ and $male \times \boldsymbol{state_a}$, so that education and state of residence are allowed to have gender-specific effects on employment by age. Finally, μ_a is a random error term. Standard errors in the second step are calculated by bootstrapping the entire estimation procedure (first and second steps together, 50 repetitions).

We also weight regressions by the number of observations used to estimate the fixed effect in the first stage, since it affects the precision of our estimates of ability. The main concern here is that the fixed effects may be estimated more precisely for men than for women, since men participate more in the labor market (and, therefore, more periods are used to estimate their fixed effects). This could potentially lead to a weaker correlation between measured ability and employment for women than for men due to attenuation bias. If this were the case, there could be a spurious positive correlation between employment and the interaction of ability with the male dummy. We use this weighting in most of our regressions, but also present robustness results without weighting.

Our coefficients of interest are θ_{1a} (the coefficient on $\hat{\gamma}^s$) and θ_{2a} (the coefficient on $male \times \hat{\gamma}^s$). A positive θ_{1a} suggests that skilled individuals are more likely to be formally employed than the less skilled (positive selection), whereas a positive θ_{2a} means that selection is more positive for men than for women at age a.

5.2. Results

Before presenting the results from our econometric analysis, we first briefly discuss our estimates of individual ability and use them graphically to illustrate our main result. Figure 5 presents the distributions of normalized individual fixed effects for each gender, while Appendix Table A.3 presents some moments of the original (non-normalized) distribution together with the probability distribution around the mean under the normalized distribution.[11] By construction, means of the normalized distributions are equal to 0 and standard deviations are equal to 1. From the figure and the numbers in Appendix Table A.3, the two normalized distributions look pretty similar, with some

[11] Notice that the mean of the original distributions are not zero because the panel is not balanced (individuals participate in the formal labor market in different periods and for a different number of periods). For the interested reader, the results for the fixed effects wage regressions are summarized in columns 3 and 4 in Appendix Table A.2.

modest differences. The female distribution has a slightly higher kurtosis than the male distribution, with a bit more density immediately below the mean, and less immediately above it and between one and two standard deviations below the mean. Since we have 443,385 observations, a Kolmogorov-Smirnov test nevertheless rejects the null hypothesis that the two distributions are identical.

Figure 6 uses the estimates of individual ability to describe the dynamics of labor market selection across ages in the 1974 cohort. In this figure, we plot the distribution of ability for individuals who were employed at four different moments (ages 21, 26, 31, and 36). The figure shows that the ability distribution of individuals employed at a point in time deteriorated for both men and women, consistent with the expansion in formal employment as individuals aged (documented in Figure 1). Individuals who held a job at age 21 had on average higher ability level, irrespective of gender, than the average individual who held a formal job at age 36. But the most important point illustrated by Figure 6 is that the shift to the left in the ability distribution was much stronger for women than for men. This difference is strong enough to be visually obvious. The mean of the ability distribution for women at age 21 was 0.20, while for men it was 0.13. By age 36, the mean for women was 0.07 and for men 0.10. A simple difference-in-differences calculated from these means suggests a relative deterioration in the mean ability of women of 10% of a standard deviation during the first 15 years of professional life. We explore this point formally by using the framework described before to analyze the relationship between ability and employment over the early lifecycle.

5.2.1. Main Results: Employment at Ages 21, 26, 31, and 36

Table 3 presents the main results from our empirical exercise. All specifications include the control variables mentioned in equation (6). Selection on ability into the labor market was positive for both genders in the 1974 cohort, and differentially higher for men at all ages, but the pattern of this relationship was not constant across ages. At age 21, soon after individuals entered the labor market, selection on ability was positive but displayed almost no gender differential. An ability level one standard deviation above the mean at that age corresponded to an advantage of 4.6 percentage points in the probability of employment for women, and of 5 percentage points for men. For the other ages considered, overall selection into the labor market became less relevant, consistent with increased formal employment for both genders as individuals aged, but differential selection across genders became stronger. For women, a one standard deviation increase in ability increased the probability of employment by 3.2 percentage points at age 26, by 1.2 percentage point at age 31, and by 1.8 percentage point by age 36. For men, the corresponding magnitudes were, respectively, 4.1, 2.7, and 2.9 percentage points. The positive and statistically significant coefficient on the interaction of the ability variable with the gender dummy indicates that positive selection was stronger for men throughout the early stages of the professional career. Comparing to the level of formal employment by gender at age 36, these numbers corresponded to increases in the probability of employment of more able individuals (one standard deviation above the mean) of 4.9 percent for women and 7.4 percent for men.

But the most important aspect of these coefficients is their differential pattern across ages. At age 21, there was little detectable differential selection across genders. By age 26, positive selection was clearly stronger for men than for women, but the difference was still quantitatively small (0.9 percentage point). After that point, the pattern of selection for women dropped very quickly with age, by more than 50 percent by age 31, and then recovered by a small amount by age 36. At the same time, differential selection across genders increased with age, so that by age 31 more able men (one standard deviation above the mean) presented an advantage in terms of probability of employment that was more than 100 percent higher than that of their female counterparts. By age 36, this relative magnitude dropped but remained 62 percent higher for high-ability men than for high-ability women.

As a result, the profile of positive selection of men dropped much more slowly than that of women: at age 36, men one standard deviation above the mean in terms of ability still displayed a probability of employment 2.9 percentage points above the mean, starting from 5 percentage points at age 21. For women during this same age interval, this number dropped from 4.6 percentage points to 1.8. This evidence suggests that the increase in formal employment observed for both genders between ages 21 and 36 documented in Figure 1 was not innocuous in terms of the relative composition of the male and female labor forces. Expansions in age-specific employment rates during the early career led to a faster deterioration of the ability pool for women than for men. This result is also consistent with the role of selection in the early career growth of the gender wage gap documented in Table 2, where it appeared to play a particularly important role during the mid-20s. The evidence suggests that, by that age, male formal workers started being selected from a relatively better portion of the ability distribution when compared to female workers. As the pool of male workers improved in relation to the pool of female workers, the unexplained part of the wage differential across genders increased.

The result that selection into employment was stronger for men than for women may be surprising, but is in line with evidence from the US documented by Bertrand et al. (2010) and Herrmann and Machado (2012). We show that, for the 1974 cohort, this same pattern was present in the context of the formal labor market in Brazil, a developing country with very different institutions.

In addition, and most importantly, we show that this pattern of differential selection across genders was exacerbated during the first years of the professional career of this cohort, thus playing a relevant role in the early career growth in the gender wage gap.

5.2.2. Robustness Exercises

The main limitation of the RAIS dataset is that it only covers formal employees. Therefore, it does not allow us to observe workers' experience in the informal sector. In the case of Brazil, informal work – encompassing self-employment and unregistered employment relationships[12] – represents an important fraction of the labor market. Omitting informal experience from our set of controls can bias the estimates if informal labor market participation is correlated with gender and with later productivity. If, for example, informal labor market experience increases productivity in the formal sector, we could be partially attributing to ability effects that are in reality due to unobserved informal experience, which might also be different across genders. On the other hand, if informal experience reduces productivity in the formal sector, as some proponents of the theory of dual labor markets have suggested (Cain, 1976), then ignoring potential experience in the informal labor market could bias our estimates of the effect of ability on formal employment downwards. As long as participation in the informal sector varies across genders and affects potential wages in the formal sector, either of these two mechanisms could interfere with our conclusions regarding differential selection across genders.

To address this concern, we estimate workers' potential experiences in self-employment and in unregistered employment from the Brazilian National Household Survey (PNAD). First, for each age-education-gender-state cell *aegs*, we compute the fraction of individuals not employed in the formal sector who work as informal employees ($pinf_{aegs}$) and as self-employed ($pself_{aegs}$). Following, 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 potential informal employment by using her *aegs* group's *pinf* and *pself* variables. For each individual-year, informal experience is summarized by two variables – *exper_inf* and *exper_self* – that indicate the cumulative sums of *pinf* and *pself* for all previous years when the individual did not appear in the RAIS dataset. These variables, therefore, capture the expected years of informal employment and self-employment given an individual's age, gender, state of residence, and previous absences from the formal labor market.

[12] An employment relationship is considered formal in Brazil when it is registered (and signed by the employer) on the worker's "labor card" (*carteira de trabalho*). An employee hired formally is entitled to social security benefits while an employer who hires formally has to pay social security and payroll taxes. If caught, an employer who hires a worker informally is subject to fines imposed by the Ministry of Labor.

Once these additional experience variables are constructed, we include them as additional controls when estimating the fixed effects Mincer equation in the first stage. Both informal and self-employed potential experiences appear as statistically significant at the 1% level when included in our 1st stage wage equations for men and for women, indicating that they are capturing relevant dimensions of individuals' labor market histories not included in the RAIS dataset. The inclusion of these variables leads to estimates of individual ability that control for potential labor market experience outside of the formal sector. We then re-estimate the specification from Table 3 using these new estimates of individual-level ability as explanatory variables. The results from this exercise are presented in Panel A from Table 4.

The qualitative pattern of the results is very similar to that from Table 3. Differential selection across genders increased between ages 21 and 31, and then receded between ages 31 and 36. Quantitatively, the only noticeable difference is that, under this specification, there was already a significant difference in selection across genders by age 21. The evidence indicates that returns to informal experience cannot account for the correlation between unobserved ability and employment estimated within our formal labor market sample. If anything, once informal experience is accounted for, differential selection across genders seems to have been even more important for the 1974 cohort than documented in Table 3.

Selection on time-varying unobservables is the second most important challenge to our empirical strategy. This could be a concern, for example, if our measure of ability based on individual fixed effects were biased due to local economic shocks for a particular group. With the extensive margin elasticity of labor supply being higher for women than for men, this might generate a gender gradient in the relationship between measured ability and employment (with, for example, women who enter the market only in periods of higher wages, and exit soon afterwards, displaying a high measured ability and a low labor supply over the entire time period).

In order to address this concern, we re-estimate our measure of individual ability correcting for selection into employment in our first stage. We apply a Heckman correction procedure when estimating the wage equation in the 1st step of our empirical exercise. The "step zero" in this correction procedure is the estimation of a probit for formal employment, where the variable excluded from the wage equation is the formal employment rate by gender-schooling-state-age cell, constructed for each year from the Brazilian National Household Survey (PNAD).[13] The goal of this variable is to capture

[13] Sector dummies and tenure are excluded from this "step zero" since we do not observe these variables for individuals without a formal job. Individual fixed effects are also excluded from "step zero," given the problems associated with

shocks to local labor demand for specific demographic groups, which might affect selection into the formal labor market and bias our estimates of individual ability.

We have a very strong "step zero" in this estimation: the F statistic of the excluded variable is of the order of 800 in the women's employment equation and 2120 in the men's equation (we estimate employment equations separately for men and women because we are also estimating separate wage equations by gender in our 1st step). The Inverse Mills Ratio is also statistically significant at 1% for both genders when included in the estimation of the wage equation in the 1st step. So our shock to local labor demand by demographic group seems indeed to do a good job in capturing forces determining time-varying selection into the formal labor market. The remainder of the empirical exercise proceeds following the same methodology outlined before.

The results from this robustness exercise are presented in Panel B from Table 4. Results remain quantitative and qualitatively very similar to those from Table 3. Time-varying selection on unobservables does not seem to be responsible for a spurious differential correlation between measured ability and employment across genders.

For the interested reader, Table 4 also presents three additional results that change the main methodological choices made in our benchmark specification. First, in Panel C, we present results of unweighted regressions in our 2nd step. Remember that we weight regressions in our main specification by the number of periods in which individuals appear in the sample (are employed formally) to avoid a weaker correlation between employment and ability for women due to the latter being measured more imprecisely for individuals appearing less in the formal labor market. In Panel D, we present results without normalizing our measure of ability (without dividing by the standard deviation of fixed effects for each gender, but still centering it on the mean). This specification allows the variance of individual ability to vary across genders and does not measure ability anymore in standard deviation units, but in monetary units.

The results show that these methodological choices do not alter the main conclusions from our empirical exercise, namely, that formal labor market selection became relatively worse for women in comparison to men between ages 21 and 31, and then recovered a bit between ages 31 and 36. In the case of the non-normalized measure of individual ability, the change in the magnitude of the estimated coefficients when compared to Table 3 is not important, since the scale of the independent variable is different from that used before. Quantitatively, according to column 3 in Panel D, for example, women

estimating fixed effects in non-linear contexts. As before, we do have individual fixed effects in the wage equation, though, so this should not be of much consequence for the results.

with unobserved ability priced 100 percent higher by the labor market had a probability of formal employment 2.4 percentage points higher than average women in the sample, while analogous men had an advantage of 5.6 percentage points.

Finally, Panel E re-estimates our entire procedure using a single wage equation for both genders in the first step (closer to the specification used in column 8 from Table 2). This specification differs from our benchmark regression in that it does not allow the coefficients on education, experience, contractual hours, tenure within the firm, firm size, sector of economic activity, and state to vary by gender. We still normalize the estimates of individual ability within genders to maintain comparability across the two distributions. The limitation of this specification, and the reason why we do not use it as our benchmark, is that it does not capture any discrimination reflected on differential returns to productive attributes across men and women, assigning it instead to ability. This specification therefore tends to introduce measurement error in the estimated ability distribution of women. For example, if women with high attachment to the labor market experience lower increases in earnings during the early career due to discrimination in promotions, this specification would attribute it to lower ability, while our benchmark specification allows the return to experience to vary by gender (possibly due to discrimination).

Panel E in Table 4, nevertheless, shows that results remain qualitatively the same when we use this specification. The overall profile of differential selection across genders is reduced by a roughly constant value – around 0.5 to 0.6 percentage point – across all ages, as should be expected from the previous discussion. Its evolution across ages, though, remains very similar, increasing monotonically up to age 31 and then falling between ages 31 and 36. It also remains statistically significant for all ages between 26 and 36.

Our main finding that selection in the 1974 cohort was typically more positive for men than for women and that this pattern was intensified as individuals aged is robust to controlling for non-formal labor market experience and for selection on time-varying unobservables, and also to variations in the key methodological choices implicit in our benchmark strategy.

5.2.3 Heterogeneity across Educational Groups

One might argue that labor markets are entirely segmented across educational groups, in which case treating all male and female workers as belonging to the same labor market would be inadequate, even more so given the differences in educational levels observed across genders. If this were the case, the analysis in principle should be conducted separately for each educational group. In order to address this concern, we replicate our main results, corresponding to the basic specifications from

Table 3, re-estimating the entire procedure – 1st and 2nd steps – separately for each of the four educational groups mentioned before: less than high school, complete high school, incomplete college, and complete college.

The results from these exercises are presented in Table 5. Appendix Figure A.1 presents the estimated distributions of individual abilities by gender and educational groups. The ability distributions by gender and schooling level differ more across genders than those estimated jointly for all educational groups (presented in Figure 5), indicating, if anything, that the assumption of equal underlying distributions of ability across genders seems more reasonable when educational groups are pooled together than when they are treated separately (which perhaps should be expected given the distinct distribution of schooling across genders; see Table 1).

Results across the first three educational levels – corresponding to Panels A, B, and C in Table 5 – are qualitatively similar to those estimated before. Differential selection across genders became relatively more positive for men within each of these groups as individuals aged. The levels and specific shapes of this profile, though, vary across groups. For the group with less than high school, differential selection was stronger than in Table 3, but its profile over time was similar to that documented before. For individuals with complete high school, women started as being more positively selected at age 21, but from then on differential selection changed favorably towards men and increased monotonically. For those with incomplete college, differential selection was more stable across ages.

For college-educated individuals, the pattern is different. Contrary to the other educational groups, absolute selection into the labor market increased almost monotonically as individuals aged. At the same time, differential selection across genders was only barely statistically significant at age 21. For later ages, differential selection was very small and not statistically significant.

The differences in the dynamics of selection across genders and educational levels reflect distinct patterns of entry and exit into the formal labor market. One possibility is that the demand for flexibility in hours and work schedules among women with lower education leads those with relatively higher ability to seek jobs outside the formal sector or to exit the labor force, while this may not happen for women with college degrees that could be able to find flexible jobs inside the formal sector. Another possibility is that high-ability women with college degrees can "afford" to work full time even after having children by hiring cheap low-skilled labor to help with household chores (e.g., nannies and housemaids). This alternative could help explain the different pattern of selection we observe for college graduates in our results for Brazil, when compared to the results found in the literature on high-skilled workers in the U.S. (Bertrand et al., 2010; Goldin, 2014). In fact, Bar et al. (2017) develop a dynamic macro model showing that the cost of low-skilled labor can be an important driver of differences in labor supply across women of different skill levels. In their model, increased inequality, by reducing wages of low-skilled workers, allows wealthier women to marketize part of the costs associated with home production and, therefore, to maintain higher labor force participation over the lifecycle. Empirical evidence from the effect of increased low-skilled immigration on the labor supply of high-skilled women also supports the relevance of this mechanism (Furtado, 2016). In the Brazilian context of very high inequality, this would suggest that differential selection across genders may not be important among the top educational group, while at the same time it may remain relevant for lower educational levels.

This is an issue that deserves further analysis but we cannot hope to address it directly with the data used in this paper. In general, Table 5 confirms that differential selection across genders in the 1974 cohort changed substantially over time, typically in favor of men, for the vast majority of the Brazilian formal labor force, even though the specific profile of this change varied by level of schooling.

6. Concluding Remarks

Interruptions in labor force participation are more common among women than among men. As a result, much of the labor economics literature focuses on differential accumulation of labor market experience as an explanation for the early career growth in the gender wage gap. We argue that differential interruptions also affect the evolution of the gender wage gap through another mechanism: selection on ability. Exit and entry into the labor market are not random, varying with unobserved ability in systematic ways. This can contribute to lifecycle changes in the unexplained portion of the wage differential across genders.

We investigate this question by using a two-step procedure and the Brazilian RAIS dataset. First, we use the panel structure of the data to reconstruct workers' labor market histories and recover a measure of unobserved individual ability. Following, we estimate regressions relating formal employment to our estimated measure of ability.

Our results show that, for the cohort born in 1974, positive selection on ability was indeed more relevant for men than for women, and that this difference grew during the early years of professional life. Male formal workers started with ability levels similar to their female counterparts, but the pool of male formal workers became relatively better during the late twenties, contributing to the increase in the gender wage gap observed in the beginning of the professional career. The age profile of these changes suggests that they are likely related to the timing of fertility decisions, though our dataset does not allow us to provide direct evidence in this direction. Our results indicate that 32% of the growth in the gender wage gap between ages 21 and 36 in this cohort is accounted for by differential changes in selection across genders.

References

Abowd, J.M., Kramarz, F., and Margolis, D.N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, 67(2), 251-333.

Adda, J., Dustmann, C., and Stevens, K. (2017). The Career Costs of Children. *Journal of Political Economy*, 125(2), 293-337.

Bar, M., Hazan, M., Leukhina, O., Weiss, D., and Zoabi, H. (2017). "Inequality and the Changing Role of Differential Fertility." Unpublished manuscript, Tel-Aviv University.

Becker, G.S. (1957). *The Economics of Discrimination*. University of Chicago Press, 178p.

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(3), 228–255.

Blau, F., and Kahn, L. (2000). Gender Differences in Pay. *Journal of Economic Perspectives*, 14(4), 75-99.

Blau, F., and Kahn, L. (2006). The US Gender Pay Gap in the 1990s: Slowing Convergence. *Industrial and Labor Relations Review*, 60(1), 45-66.

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.

Blau, F., and Kahn, L. (2016). "The Gender Wage Gap: Extent, Trends, and Explanations." IZA DP No. 9656.

Cain, G. (1976). The Challenge of Segmented Labor Market Theories to Orthodox Theory: A Survey. *Journal of Economic Literature*, 14(4), 1215-1257.

Cahuc, P., and Zyberberg, A. (2004). *Labor Economics*. Cambridge: MIT Press.

Card, D., Cardoso, A.R., and Kline, P. (2016). Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women. *Quarterly Journal of Economics*, 131(2), 633-686.

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(3), 417-441.

Corseuil, C.H., Da Silva, A., Dias, R., and Maciente, A. (2010). "Consistência das Bases de Dados do Ministério do Trabalho." Unpublished manuscript, IPEA-Rio.

Corseuil, C.H., Foguel, M., Gonzaga, G., and Ribeiro, E. (2013). "Youth Labor Market in Brazil through the Lens of the Flow Approach." Paper presented at the XLI Encontro Brasileiro de Econometria.

Cruces, G., Ham, A., and Viollaz, M. (2012). "Scarring effects of youth unemployment and informality: Evidence from Brazil." Unpublished manuscript.

Fernandes, M. (2013). "Ensaios em Microeconomia Aplicada." Unpublished PhD Dissertation, Department of Economics, PUC-Rio.

Furtado, D. (2016). Fertility Responses of High-Skilled Native Women to Immigrant Inflows. *Demography*, 53(1), 27-53.

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.

Goldin, C. (2014). A Grand Gender Convergence: Its Last Chapter. *American Economic Review*, 104(4), 1091–1119.

Goldin, C., Kerr, S.P., Olivetti C., and Barth, E. (2017). The Expanding Gender Earnings Gap: Evidence from the LEHD-2000 Census. *American Economic Review: Papers and Proceedings*, 107(5), 110-114.

Heckman, J.J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, 42(4), 679-694.

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

Li, I., and Miller, P. (2012). "Gender Discrimination in the Australian Graduate Labour Market." IZA Discussion Paper 6595.

Machado, C. (2017). Unobserved Selection, Heterogeneity and the Gender Wage Gap. *Journal of Applied Econometrics*, forthcoming.

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(2), S76-S108.

Mulligan, C., and Rubinstein, Y. (2008). Selection, Investment, and Women's Relative Wages over Time. *Quarterly Journal of Economics*, 123(3), 1061-1110.

Ñopo, H. (2012). *New Century, Old Disparities: Gender and Ethnic Earnings Gaps in Latin America and the Caribbean*. Washington D.C.: Inter-American Development Bank and World Bank.

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(1), 119-140.

Appendix

A1. Data Issues in the RAIS and Correction Algorithms

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 should 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. When 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 could 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 whom 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.

The final dataset contains 443,392 individuals, 44.1% of which (195,331) are women. The correction of the education variable, in particular, is not possible for many individuals, reducing the sample size by 18.7%. It is also worth noting that data inconsistencies seem to be more common for men, since the percentage of women increases with the data correction and sample selection procedures.

Table 1: Descriptive Statistics, RAIS and Census Datasets, 1995-2010, 1974 Cohort, Brazil

Variables	All	Women	Men
Wage	1361.65	1226.19	1455.17
	(1901.54)	(1737.50)	(2001.66)
lwage	6.855	6.758	6.922
	(0.75)	(0.73)	(0.75)
Age	29.08	29.31	28.92
	(4.57)	(4.55)	(4.57)
Hours	41.74	40.54	42.57
	(5.53)	(6.86)	(4.19)
N Observations (Worker x Year)	3,639,101	41%	59%
Schooling (RAIS):			
Less than high school	40.1%	30.7%	47.4%
Complete high school	40.9%	44.5%	38.1%
Incomplete college	3.8%	4.4%	3.3%
Complete college	15.2%	20.4%	11.2%
N Observations (Workers)	443,385	44%	56%
Schooling (Census, All population):			
Less than high school	58.0%	54.7%	61.5%
Complete high school	26.9%	28.3%	25.5%
Incomplete college	2.5%	2.3%	2.7%
Complete college	12.6%	14.8%	10.3%
N Observations	254,762	50%	50%
Schooling (Census, Formal Workers):			
Less than high school	46.6%	34.8%	53.9%
Complete high school	33.9%	37.2%	31.9%
Incomplete college	3.1%	3.5%	2.9%
Complete college	16.3%	24.4%	11.3%
N Observations	79,380	37%	63%

Notes : Standard errors in parenthesis. Variable *Wage* is average monthly earnings in Brazilian reais of December 2010. Variable *lwage* is its natural logarithm. Variable *hours* is contracted weekly working hours, and *age* is the individual's age. All calculations use the full sample.

Table 2: Evolution of the Gender Wage Gap by Age, OLS and Fixed Effects, RAIS Dataset, 1995-2010, 1974 Cohort, Brazil

	OLS without Controls for Labor Market History		OLS with Controls for Labor Market History		Cumulative Δ in W Gap Explained by Labor Market		FE Model, Controlling for LM History	Cumulative Δ in W Gap Explained by Selection	
Age	Gender W gap by age	Cumulative Δ in W gap by age	Gender W gap by age	Cumulative Δ in W gap by age	Level	%	Cumulative Δ in W gap by age	Level	%
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
21	0.131		0.127						
22	0.147	0.0160	0.140	0.0130	0.0030	18.6%	0.0023	0.0107	82.6%
23	0.172	0.0410	0.165	0.0378	0.0033	7.9%	0.0145	0.0232	61.5%
24	0.185	0.0541	0.175	0.0484	0.0057	10.6%	0.0161	0.0323	66.7%
25	0.200	0.0690	0.189	0.0622	0.0068	9.8%	0.0166	0.0456	73.3%
26	0.221	0.0907	0.209	0.0825	0.0083	9.1%	0.0318	0.0506	61.4%
27	0.238	0.1070	0.224	0.0968	0.0102	9.6%	0.0451	0.0517	53.4%
28	0.247	0.1161	0.231	0.1038	0.0123	10.6%	0.0480	0.0558	53.8%
29	0.267	0.1365	0.250	0.1235	0.0130	9.5%	0.0612	0.0623	50.4%
30	0.284	0.1538	0.266	0.1395	0.0143	9.3%	0.0783	0.0612	43.9%
31	0.292	0.1611	0.272	0.1446	0.0165	10.2%	0.0831	0.0615	42.6%
32	0.298	0.1676	0.278	0.1509	0.0167	10.0%	0.0881	0.0628	41.6%
33	0.313	0.1826	0.293	0.1659	0.0166	9.1%	0.1016	0.0643	38.8%
34	0.326	0.1952	0.305	0.1779	0.0173	8.9%	0.1153	0.0626	35.2%
35	0.327	0.1959	0.305	0.1782	0.0177	9.0%	0.1164	0.0618	34.7%
36	0.333	0.2027	0.316	0.1893	0.0133	6.6%	0.1296	0.0597	31.6%
Exper. Controls	No	No	Yes	Yes			Yes		
F-Stat (Exper. Vars.)			1301.73				1278.18		

Notes : Columns 2 and 4 present the level of the gender wage gap by age estimated from OLS wage regressions (interactions of age dummies and a male dummy), controlling and not controlling for previous formal labor market history (a large set of non-parametric controls for experience, described in detail in the text). Columns 3 and 5 present the respective cumulative change in the gender wage gap by age corresponding to columns 2 and 4, while column 8 presents the cumulative change in the gender wage gap by age estimated from an equation with individual fixed effects (which also controls for previous formal labor market history). All wage equations have the log of monthly real earnings as dependent variable and include as additional controls four education dummies, tenure, weekly working hours, age (year) dummies, state dummies, aggregate sector dummies, and firm size dummies (see text). Column 6 shows the difference in levels in the cumulative growth in the gender wage gap between columns 3 and 5, and column 7 shows this difference as a fraction of column 3. Similarly, column 9 shows the difference in levels in the cumulative growth in the gender wage gap between columns 5 and 8, and column 10 shows this difference as a fraction of column 5.

Table 3: Employment at Different Ages, Linear Probability Models, RAIS Dataset, 1995-2010, 1974 Cohort, Brazil

	Dependent Variable: Employment Dummy (Full-Time Most-Year)			
	Age 21	Age 26	Age 31	Age 36
	(1)	(2)	(3)	(4)
All workers				
Ability (f.e.)	0.0456***	0.0321***	0.0116***	0.0181***
	(0.00141)	(0.00177)	(0.00178)	(0.00159)
Ability × Male	0.00444*	0.00898***	0.0157***	0.0112***
	(0.00227)	(0.00192)	(0.00207)	(0.00232)
N Observations	443,385	443,385	443,385	443,385
R-squared	0.054	0.032	0.024	0.020

Notes : Bootstrapped standard errors (50 repetitions) in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is a dummy equal to one when the individual worked full-time most-year at each age-gender. FE (ability) is the normalized version of worker fixed effects. All regressions control for a gender dummy *male* , four education groups (see text), state dummies, and for the interaction of *male* with the four education dummies and state dummies. Regressions use the full sample of individuals born in 1974 with filters described in the text.

Table 4: Robustness Analyses, Employment at Different Ages, Linear Probability Models, RAIS, 1995-2010, 1974 Cohort, Brazil

	Dependent Variable: Employment Dummy (Full-Time Most-Year)			
	Age 21	Age 26	Age31	Age 36
	(1)	(2)	(3)	(4)
A. Controlling for Potential Informal Experience				
Ability (f.e.)	0.0399***	0.0288***	0.00978***	0.0168***
	(0.002)	(0.002)	(0.002)	(0.002)
Ability × Male	0.0123***	0.0143***	0.0171***	0.00953***
	(0.002)	(0.002)	(0.002)	(0.002)
N Observations	398,279	398,279	398,279	398,279
R-squared	0.047	0.032	0.024	0.019
B. Correcting for Selection (3-Stage Estimation)				
Ability (f.e.)	0.0446***	0.0308***	0.0111***	0.0195***
	(0.001)	(0.002)	(0.002)	(0.002)
Ability × Male	0.00529**	0.0105***	0.0161***	0.0111***
	(0.002)	(0.002)	(0.002)	(0.002)
N Observations	443,385	443,385	443,385	443,385
R-squared	0.053	0.032	0.024	0.020
C. Unweighted				
Ability (f.e.)	0.0455***	0.0346***	0.00396	0.00333
	(0.003)	(0.004)	(0.004)	(0.003)
Ability × Male	0.000776	0.00751*	0.0151***	0.0143***
	(0.005)	(0.004)	(0.004)	(0.005)
N Observations	443,385	443,385	443,385	443,385
R-squared	0.049	0.037	0.034	0.030
D. Centered but Not Normalyzed Measure of Ability				
Ability (f.e.)	0.0939***	0.0662***	0.0239***	0.0374***
	(0.001)	(0.001)	(0.001)	(0.001)
Ability × Male	0.00798***	0.0175***	0.0317***	0.0223***
	(0.002)	(0.002)	(0.002)	(0.002)
N Observations	443,385	443,385	443,385	443,385
R-squared	0.053	0.032	0.024	0.020
E. One Regression in First Step				
Ability (f.e.)	0.0462***	0.0348***	0.0152***	0.0214***
	(0.001)	(0.002)	(0.002)	(0.002)
Ability × Male	0.00293	0.00390**	0.00940***	0.00573**
	(0.002)	(0.002)	(0.002)	(0.002)
N Observations	443,385	443,385	443,385	443,385
R-squared	0.053	0.032	0.024	0.020

Notes: Bootstrapped standard errors (50 repetitions) in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is a dummy equal to one when the individual worked full-time most-year at each age-gender. FE (ability) in panels A-C and E is the normalized version of worker fixed effects and the centered (not normalized) version of fixed effects in panel D . All regressions control for a gender dummy *male* , four education groups (see text), state dummies, and for the interaction of *male* with the four education dummies and state dummies. Regressions use the full sample of individuals born in 1974 with filters described in the text.

Table 5: Employment at Different Ages by Level of Schooling, Linear Probability Models, RAIS Dataset, 1995-2010, 1974 Cohort, Brazil

	Dependent Variable: Employment Dummy (Full-Time Most-Year)			
	Age 21	Age 26	Age 31	Age 36
	(1)	(2)	(3)	(4)
A. Less than High-School				
Ability (f.e.)	0.0523***	0.0399***	0.0113***	0.00217
	(0.00290)	(0.00276)	(0.00316)	(0.00292)
Ability × Male	-0.00155	0.0115***	0.0222***	0.0197***
	(0.00308)	(0.00379)	(0.00388)	(0.00329)
N Observations	177,582	177,582	177,582	177,582
R-squared	0.048	0.049	0.025	0.012
B. Complete High School				
Ability (f.e.)	0.0417***	0.0310***	0.00223	0.00178
	(0.00215)	(0.00278)	(0.00236)	(0.00226)
Ability × Male	-0.00910***	-0.00147	0.0170***	0.0262***
	(0.00270)	(0.00306)	(0.00319)	(0.00289)
N Observations	181,365	181,365	181,365	181,365
R-squared	0.048	0.033	0.025	0.013
C. Incomplete College				
Ability (f.e.)	0.0297***	0.0107	-0.00622	-0.00515
	(0.00536)	(0.00696)	(0.00673)	(0.00501)
Ability × Male	0.0113	0.0186*	0.0191**	0.0182**
	(0.00825)	(0.01051)	(0.00832)	(0.00772)
N Observations	16,904	16,904	16,904	16,904
R-squared	0.062	0.033	0.020	0.012
D. Complete College				
Ability (f.e.)	0.0176***	0.0351***	0.0306***	0.0413***
	(0.00263)	(0.00332)	(0.00330)	(0.00256)
Ability × Male	0.00760*	0.00637	0.00413	-0.00262
	(0.00408)	(0.00532)	(0.00521)	(0.00379)
N Observations	67,534	67,534	67,534	67,534
R-squared	0.126	0.043	0.035	0.034

Notes : Bootstrapped standard errors (50 repetitions) in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is a dummy equal to one when the individual worked full-time most-year at each age-gender. FE (ability) is the normalized version of worker fixed effects. All regressions control for a gender dummy *male* , four education groups (see text), state dummies, and for the interaction of *male* with the four education dummies and state dummies. Panels A, B, C and D present results for workers with less than complete high-school, complete high school, incomplete college and college education, respectively. All regressions use, for each education group, the full sample of individuals born in 1974 with filters described in the text.

Appendix Table A1 - Data Filters and Sample Size, RAIS Data between 1995 and 2010, Number of Observations Refer to Individuals

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

Notes : The initial dataset is a 30% random sample of workers born in 1974 who appear in the RAIS dataset at some point between 1995 and 2010.

Appendix Table A2: Fixed-Effects Mincerian Equations, RAIS Dataset, 1995-2010, 1974 Cohort, Brazil

	Dependent Variable: Log Wages			
	Women	Men	Women	Men
	(1)	(2)	(3)	(4)
Complete High School	0.00234*	0.00647***	0.00832***	0.00506***
	(0.00134)	(0.00109)	(0.00134)	(0.00108)
Incomplete College	0.0789***	0.0906***	0.0787***	0.0839***
	(0.00237)	(0.00243)	(0.00234)	(0.00240)
Complete College	0.316***	0.384***	0.297***	0.353***
	(0.00222)	(0.00236)	(0.00221)	(0.00236)
Tenure	0.00168***	0.00169***	0.000167***	0.000347***
	(1.23e-05)	(1.04e-05)	(1.51e-05)	(1.26e-05)
Hours	0.00694***	0.00473***	0.00603***	0.00358***
	(7.49e-05)	(0.000102)	(8.03e-05)	(0.000108)
Experience controls	No	No	Yes	Yes
F-Stat.: Experience Controls			564.24	741.32
N Observations	1,486,377	2,152,724	1,486,377	2,152,724
N Individuals	195,332	248,053	195,332	248,053
R-squared	0.274	0.288	0.293	0.304

Notes: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1 . All regressions use the full sample of individuals born in 1974 and control for tenure, hours, year dummies, state dummies, firm size dummies, aggregate sector dummies, and worker fixed effects.

Appendix Table A3 - Moments of the Distributions of Fixed-Effects by Gender, RAIS Data between 1995 and 2010, Number of Observations Refer to Individuals

	Original (Non-normalized) Distribution Moments					
	N. Obs.		Mean	Std. Dev.	Min.	Max.
Females	195,332		-0.06	0.49	-1.93	3.34
Males	248,053		-0.04	0.49	-2.06	4.14

	Normalized Probability Distribution around the Mean (%)					
	x < -2sd	-2sd < x < -1sd	-1sd < x < mean	mean < x < 1sd	1sd < x < 2sd	2sd < x
Females	0.46	9.89	49.66	26.10	8.98	4.90
Males	0.38	11.58	45.85	28.31	9.17	4.71

Notes : Fixed effects computed for the 1974 cohort, using RAIS data between 1995 and 2010. Calculated based on the methodology described in section 5.1.

Figure 1: Formal Employment Rate by Gender - Data from RAIS - 1974 Cohort, Brazil, 1995-2010

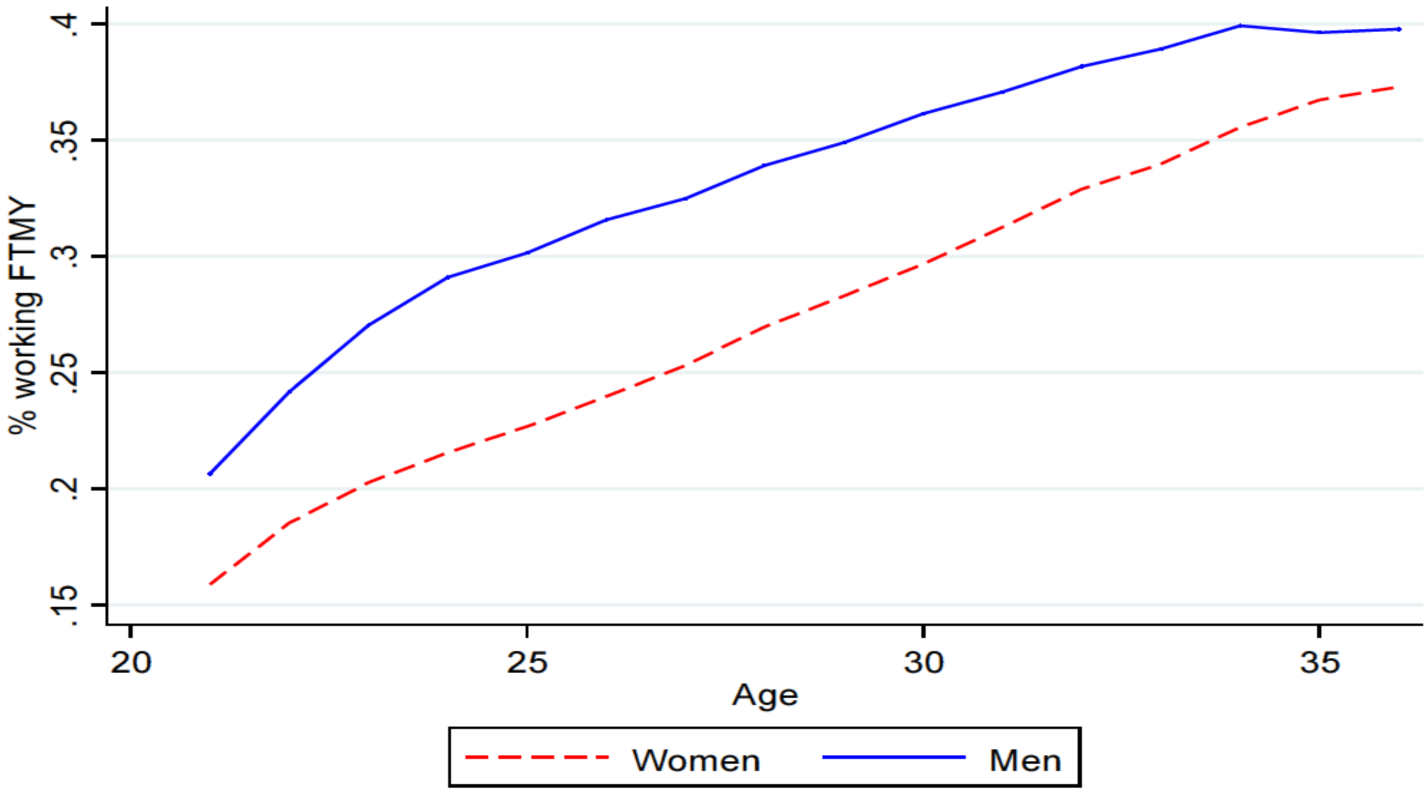


Figure 2: Survival in Formal Employment for Individuals Employed at age 21 by Gender - Data from RAIS - 1974 Cohort, Brazil, 1995-2010

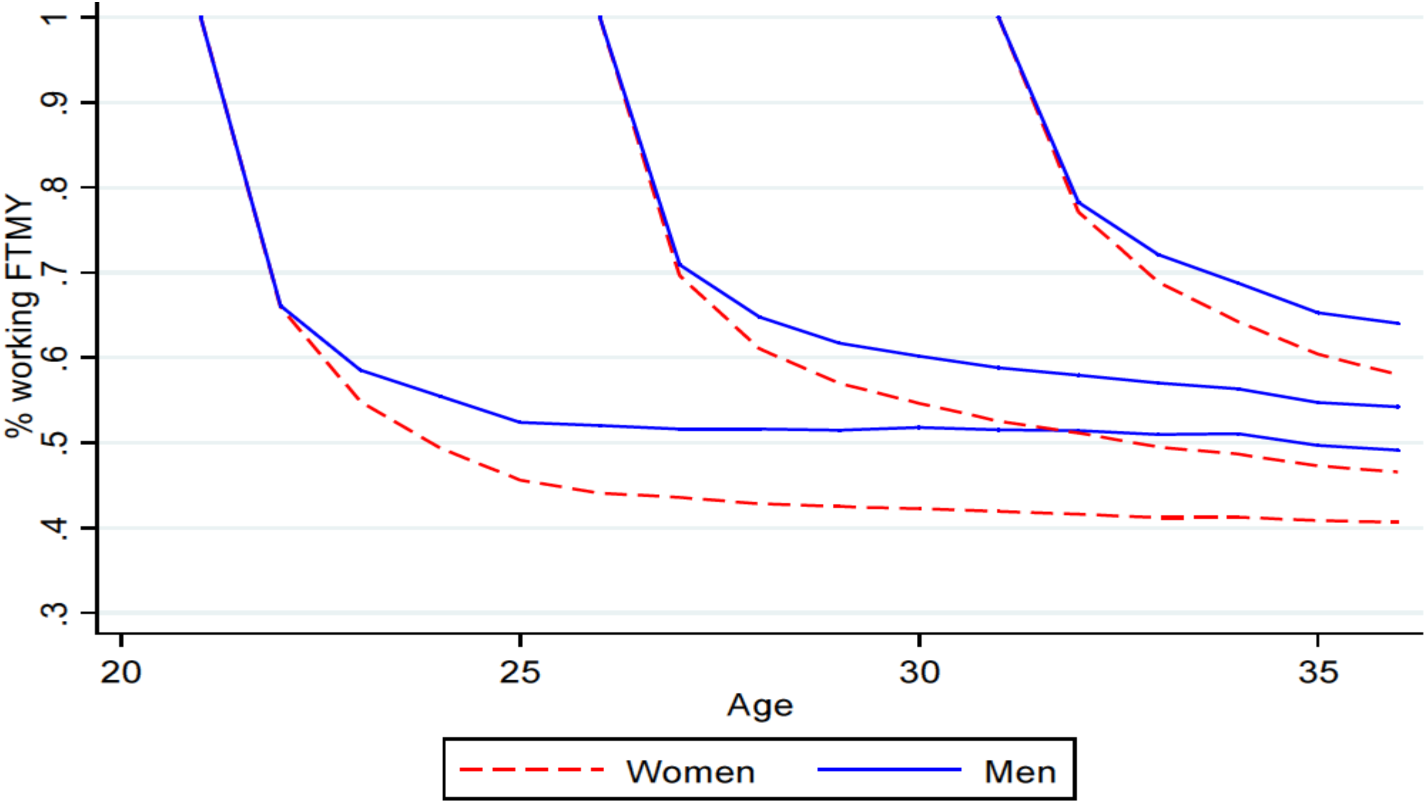


Figure 3: Returns to Experience - Wage Gain from Continuous Employment after Age 21 - Estimates from RAIS - 1974 Cohort, Brazil, 1995-2010

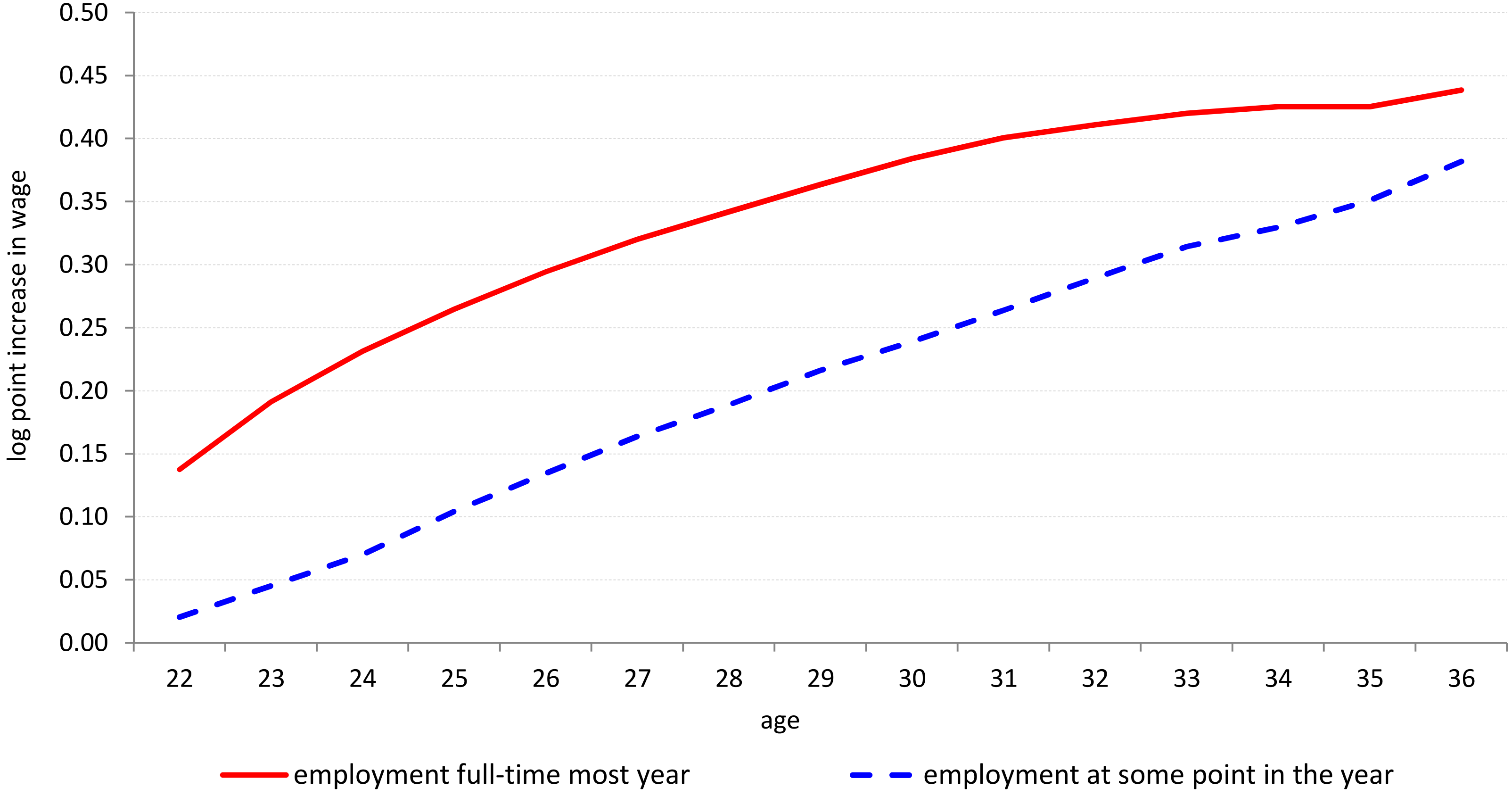


Figure 4: Cumulative Growth in the Gender Wage Gap by Age - Controlling and Not Controlling for Previous Labor Market History and Selection - Estimates from RAIS - 1974 Cohort, Brazil, 1995-2010

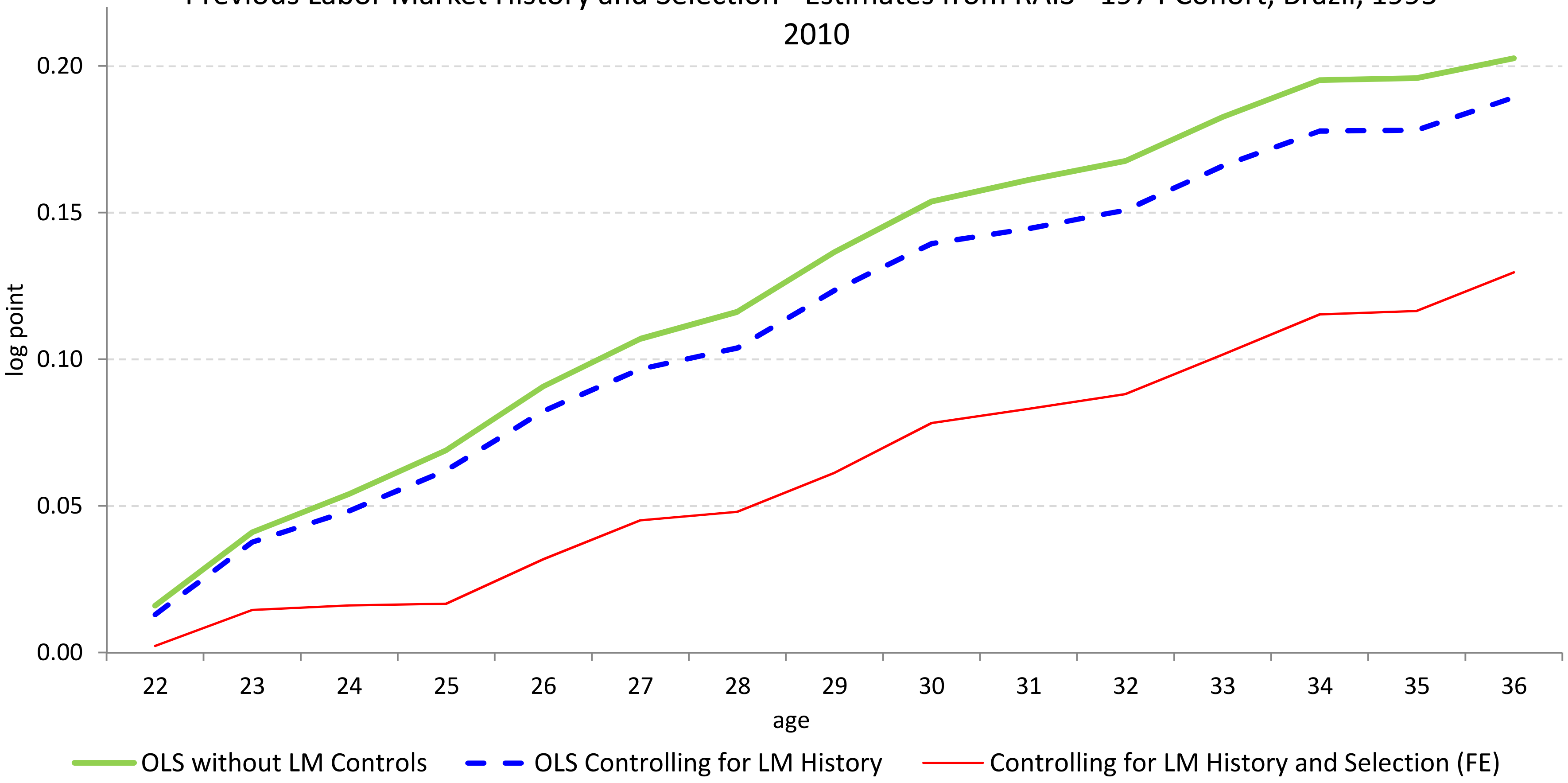
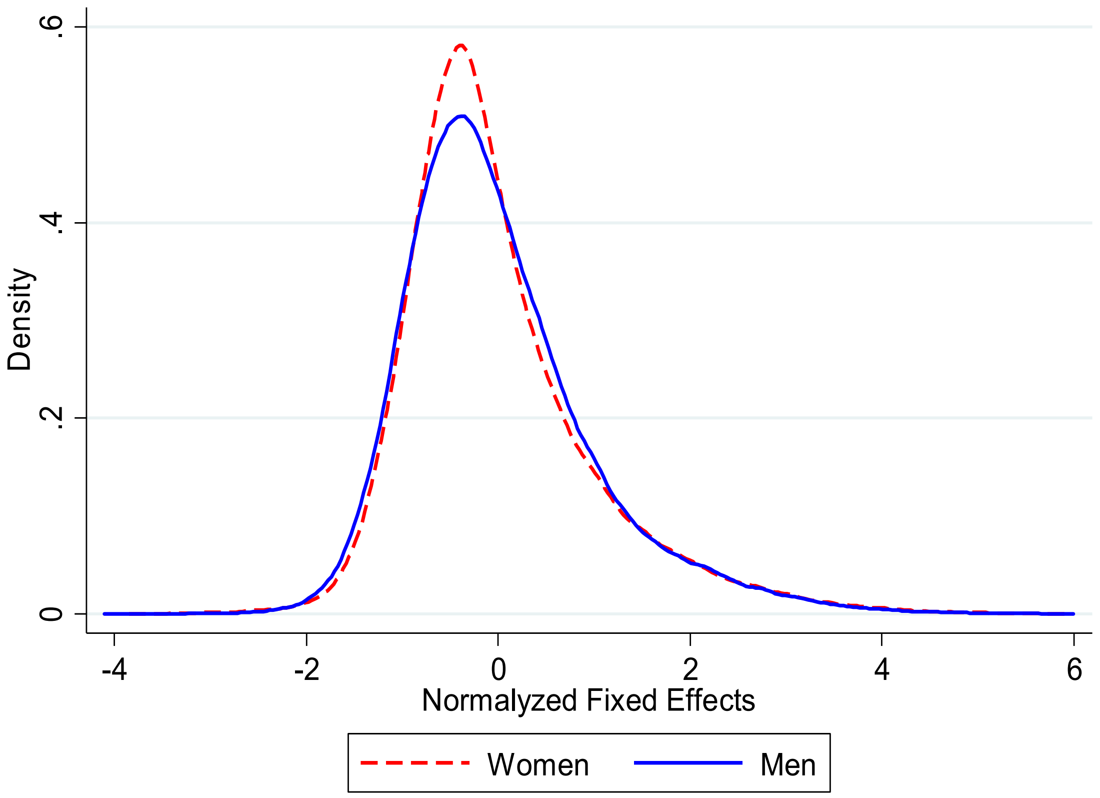
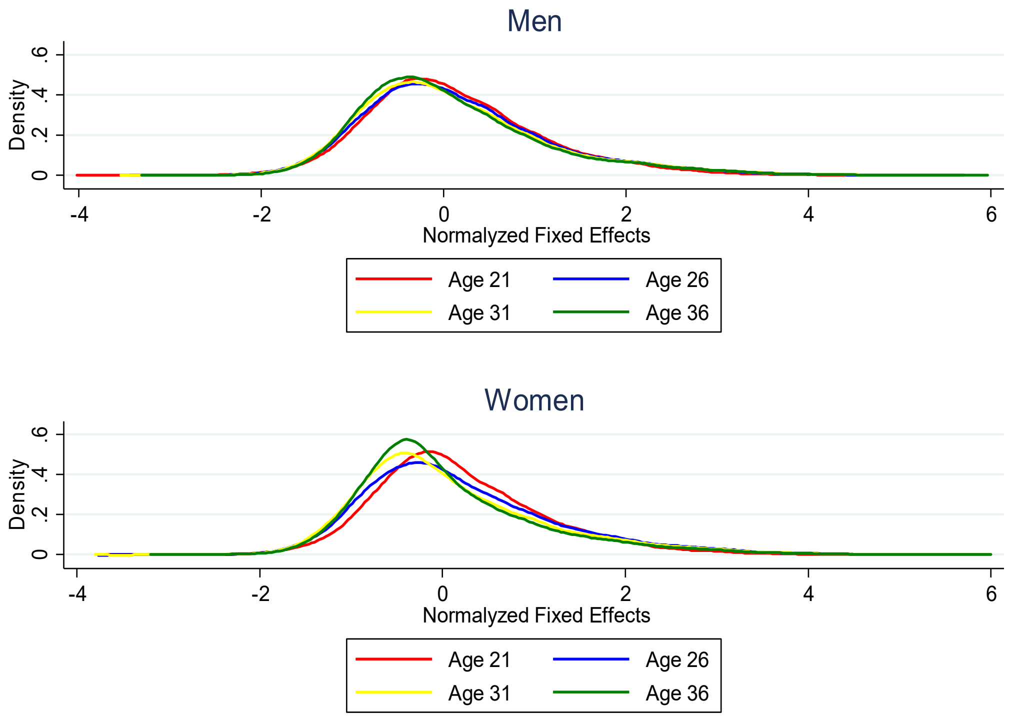


Figure 5: Estimated Distributions of Normalized Ability by Gender - Estimates Based on Equation 5 and Data from RAIS - 1974 Cohort, Brazil, 1995-2010



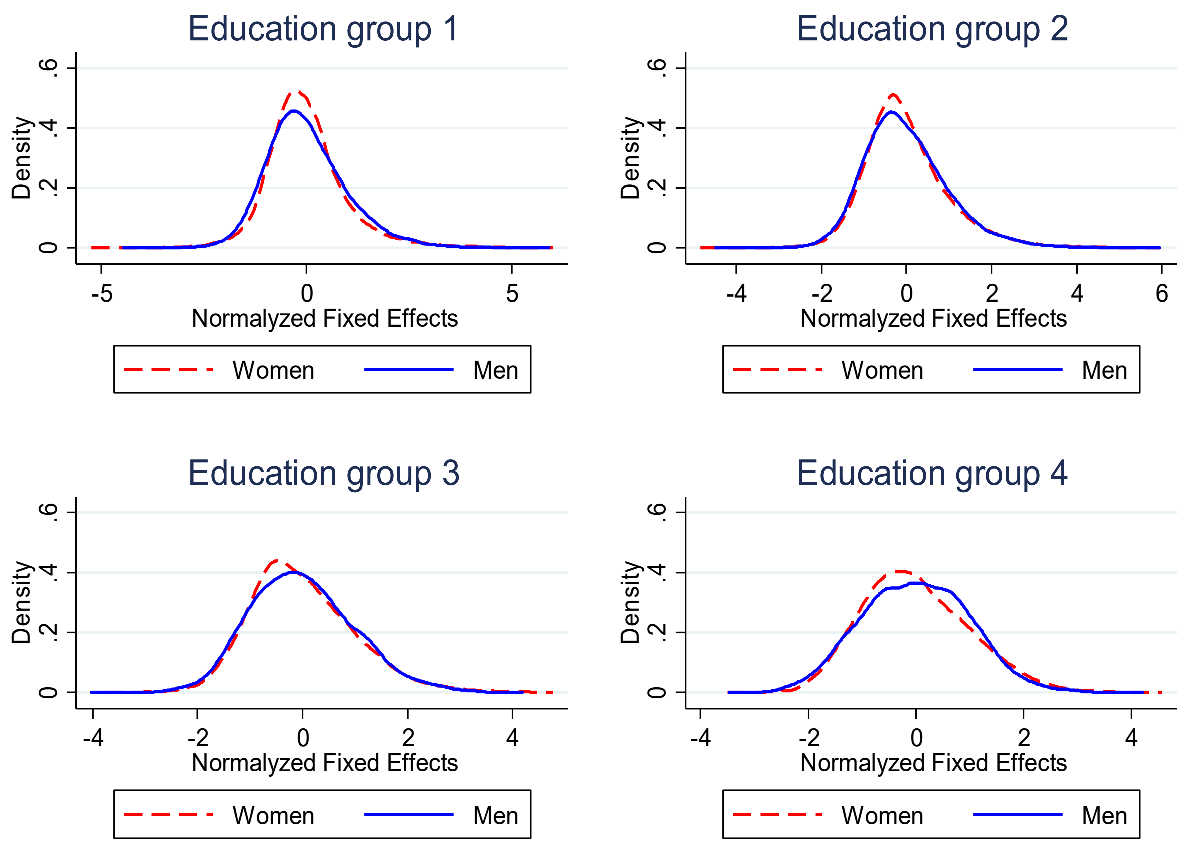
Note: Kernel density estimation uses the Epanechnikov kernel.

Figure 6: Estimated Distributions of Normalized Ability by Gender and Age - Estimates Based on Equation 5 and Data from RAIS - 1974 Cohort, Brazil, 1995-2010



Note: Kernel density estimation uses the Epanechnikov kernel.

Appendix Figure A1: Estimated Distributions of Normalized Ability by Gender and Educational Group - Estimates Based on Equation 5 and Data from RAIS - 1974 Cohort, Brazil, 1995-2010


Education group 1
Density
Normalyzed Fixed Effects
Women
Men
Education group 2
Density
Normalyzed Fixed Effects
Women
Men
Education group 3
Density
Normalyzed Fixed Effects
Women
Men
Education group 4
Density
Normalyzed Fixed Effects
Women
Men