August 2012

Essays on Informal Labor Markets

Javier Cano Urbina
The University of Western Ontario

Supervisor
Audra J. Bowlus and Lance J. Lochner
The University of Western Ontario

Graduate Program in Economics

A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

ESSAYS ON INFORMAL LABOR MARKETS

(Spine title: Essays on Informal Labor Markets)
(Thesis format: Integrated Article)

by

Javier Cano Urbina
Graduate Program in Economics

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES
THE UNIVERSITY OF WESTERN ONTARIO
LONDON, CANADA
MAY 2012

THE UNIVERSITY OF WESTERN ONTARIO
SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisors	Examiners
Dr. Audra J. Bowlus	Dr. Pedro Carneiro
Dr. Lance J. Lochner	Dr. Timothy G. Conley
Supervisory Committee	Dr. Salvador Navarro
Dr. Youngki Shin	Dr. Paul-Philippe Paré

The thesis by

Javier Cano Urbina

entitled

Essays on Informal Labor Markets

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Date _____ _____
Chair of the Thesis Examining Board

Abstract

This thesis consists of three related papers. The first paper examines whether informal sector jobs are a source of training for young less-educated workers. Controlling for worker and job characteristics, it is found that in the early years of workers' careers in Mexico, wage growth in the informal sector is higher than in the formal sector. This result is consistent with general human capital investment on-the-job if the informal labor market is more competitive than the formal labor market due to frictions generated by labor regulations. These results motivate a deeper analysis of the informal labor market which is presented in the second paper.

The second paper examines two roles that informal sector jobs play in the early stages of a worker's career: informal jobs may (*i*) provide the opportunity to accumulate skills, and (*ii*) act as a screening device that enables employers to learn a worker's ability. This paper develops a matching model of the informal and formal sectors that can accommodate both roles. Implied hazard rates from informal to formal sectors as a function of tenure are shown to differ depending on whether the role of informal sector jobs is human capital accumulation or screening. Using the ENOE, a longitudinal employment survey from Mexico, hazard functions are estimated for less-educated workers. The estimated hazard functions suggest the informal sector plays an important role by screening less-educated workers in the early stages of their careers. The estimation results also imply that employers would only learn the ability of 14% of their workers after one month of employment. This finding suggests that employers'

capacity to select workers is limited in government employment programs requiring employers to provide permanent positions to a predetermined fraction of workers after a short period of time.

The duration data used for estimation in the second paper is obtained from the stock of individuals employed in the informal sector at a given point in time. It is known that duration data obtained from a given stock of individuals can fail to observe those with relatively short spells. Accounting for this sample bias requires constructing a conditional likelihood function, which in turn requires knowledge of the exact starting times of each spell. Unfortunately, it is common in duration data to have coarse measures for starting times, complicating the resolution of sampling bias. The third paper investigates several alternatives for overcoming coarseness by imputing interval-censored starting times and performing a Monte Carlo analysis. The results indicate that imputed interval midpoints outperform the alternatives.

Acknowledgements

First and foremost, I want to thank my two supervisors Audra Bowlus and Lance Lochner. Their numerous suggestions and insightful comments are the pillars of this thesis. They have my greatest appreciation and admiration, and I have been fortunate to have them as my mentors and role models.

I also want to thank Youngki Shin for taking the time to look at my work and providing helpful suggestions for improvement. Many of the ideas that motivated the papers in this thesis were the result of discussions with Aldo Colussi, to whom I am deeply grateful. Finally, I want to thank Todd Stinebrickner, first, for giving me the opportunity to fulfill my goals at Western, and second, for helping me to better convey my message.

I want to extend a very special thank to Yvonne Adams. She has been a good friend and a great support. Not only is she extremely efficient in what she does, but also, she goes out of her way to help students beyond what her job requires, making everyone of us feel special.

Throughout this project, I benefited from the support of my friends Daniel Montanera and Deanna Walker. I thank them for listening to my ideas, motivating me, and helping me clarify my thoughts. I have also benefited from the interaction with many other friends, helping me to move ahead in my work and making the graduate experience at Western more enjoyable. I extend my gratitude to Jon Rosborough, George (Ye) Jia, Jacob Wibe, Michael McCausland, Douwere Grekou, Shiddarta Vásquez Córdoba, Philippe Belley, David

Fieldhouse, Andrew Agopsowicz, Nick Bedard, Masashi Miyairi, Utku Suleymanoglu, Kai Xu, and William Pouliot. I only hope to have been as good of friend to you as you have been to me.

I am also grateful to have had the opportunity to interact with numerous faculty members of the Western community and getting suggestions from them on how to proceed or make my discussion clearer. It has been a honor to me to have met professors Tim Conley, Igor Livshits, Benjamin Lester, Salvador Navarro, Greg Pavlov, Chris Robinson, and Al Slivinski. I want to thank them all for listening to me and making me feel like a colleague.

Certainly having the support from the administrative staff helped in making my stay at Western more enjoyable. I want to thank Jane McAndrew, Jennifer Hope, Sharon Phillips, and Debra Merrifield for being good friends and helping me make my life less complicated.

Not being a native English speaker, I want to acknowledge all the help that I received from the members of the Writing Support Centre at the University of Western Ontario. I want to thank Ryan Robb, Thila Varghese, Derek Lattimer, and Emily Kress for their many suggestions over these years.

Finally, I want to thank the Consejo Nacional de Ciencia y Tecnología, CONACYT, for their financial support, and Lance Lochner for the opportunity to work for him as a Research Assistant, through which I learned so much and supported my studies here in Canada.

A mis padres: Lupita y Daniel.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Introduction

> The term *informality* means different things to different people, but almost always bad things
>
> Maloney and Saavedra-Chanduvi (2007)

This dissertation is composed of three related papers. The first two papers, presented in Chapters 2 and 3, study the role of informal jobs over the career of less-educated workers. The third paper, presented in Chapter 4, studies the properties of the estimators used in Chapter 3.

An informal job is a job that does not comply with labor regulations. As such, these jobs constitute what is typically known as the informal sector. Its counterpart, the formal sector, is composed of jobs that comply with labor regulations. These regulations, such as minimum wage, health insurance, severance payment, or retirement pension, are mainly intended to protect workers,

and it is commonly argued that observance of these mandates is of great significance in developing countries to ensure social justice for workers (Berg and Kucera, 2008) and to protect them against the forces of reallocation in the labor market (Inter-American Development Bank, 2003). On the other hand, labor regulations raise labor costs, and so there is an incentive for employers not to comply. Similarly, individual workers may prefer a more direct compensation as opposed to the indirect protection offered by regulations. Furthermore, it is usually the case that developing countries have low levels of enforcement of these regulations. As a result, a mass of informal sector jobs emerge; and this jobs become the main source of employment for certain groups of the population, such as the group of young less-educated workers.

Given the importance of informal sector jobs for the employment of young less-educated workers, it is natural to try to learn more about the work experiences of this group in the informal sector. Chapter 2 provides an initial step to better understand the effects of informal jobs in the careers of less-educated workers. Evidence presented in this chapter indicates that for the group of less-educated workers, wage growth is higher in the informal than in the formal sector, once controlling for worker and job observable characteristics. This result is consistent with theories of human capital accumulation for the following reasons. First, the labor market in the informal sector is more competitive than the labor market in the formal sector. Second, in any competitive labor market, workers bear the cost of training and get wage returns (Becker, 1993).

Third, in a frictional labor market, employers benefit from workers' training and are willing to sponsor at least part of the cost of training (Acemoglu and Pischke, 1999). As a result, one might expect informal sector workers to have faster wage growth than formal sector workers. These results indicate that it is possible that informal sector jobs represent a source of training for young-less educated workers in Mexico.

Chapter 3 further explores the results of Chapter 2. The goal of this chapter is to determine if working in the informal sector can improve the career prospects of less-educated workers. To that end, this chapter considers two mechanisms through which informal jobs may positively affect the careers of less-educated workers. The first mechanism has informal sector jobs providing training opportunities for young less-educated workers. The second mechanism has informal sector jobs helping to resolve an information problem about the initially unobserved skills of young less-educated workers. These two mechanisms are separately incorporated into a matching model and testable implications are derived. The matching model developed in this paper follows the model proposed by Albrecht, Navarro, and Vroman (2006, 2009). The testable implications are based on the shapes of the hazard function from the informal to the formal sector. Each of the two proposed mechanisms implies different shapes for this function. A flexible hazard function is estimated using data from Mexico, and the estimated hazard is consistent with the implications of the second mechanism in which informal sector jobs have the function of a

screening device that helps to resolve the information problem about the initially unknown skills of young less-educated workers. It is important to mention that this result does not rule out the possibility that informal jobs also provide training to young less-educated workers.

The estimation of the hazard function in Chapter 3 required employment duration data in which is necessary to know both the lengths of the job spells and the starting dates of these spells. However, in the duration data available for estimation, some of the job spells have a coarser measure of the starting dates. In particular, for some job spells, the starting date is only known within a year, and so the starting date of the job spell is only known to be contained in an interval. As a consequence, the estimation procedure suggested in the literature (e.g. Klein and Moeschberger, 1997; Wooldridge, 2002) cannot be directly implemented. Chapter 4 explores the finite sample properties of estimates of the hazard function using the estimation procedure typically suggested in the literature, but replacing the missing starting dates of the job spells with imputed starting dates. Three imputation methods are proposed, using: (i) the lower bound of the interval, (ii) the midpoint of the interval, and (iii) the upper bound of the interval containing the starting date. A Monte Carlo analysis is performed, and the results indicate that using the midpoint of the interval outperforms the alternatives, particularly when the duration data has features similar to those of the duration data used for estimation in Chapter 3.

1.1 Bibliography

ACEMOGLU, D. AND J.-S. PISCHKE (1999): "The Structure of Wages and Investment in General Training," *The Journal of Political Economy*, 107, 539–572.

ALBRECHT, J., L. NAVARRO, AND S. VROMAN (2006): "The Effects of Labor Market Policies in an Economy with an Informal Sector," Discussion Paper IZA DP No. 2141, The Institute for the Study of Labor (IZA).

——— (2009): "The Effects of Labour Market Policies in an Economy with an Informal Sector," *The Economic Journal*, 119, 1105–1129.

BECKER, G. S. (1993): *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, Chicago: The University of Chicago Press (for NBER), 3d ed.

BERG, J. AND D. KUCERA (2008): in *In defence of labour market institutions : cultivating justice in the developing world*, ed. by J. Berg and D. Kucera, Basingstoke, Hampshire ; New York: Palgrave Macmillan, chap. 1, 1–8.

INTER-AMERICAN DEVELOPMENT BANK (2003): *Good Jobs Wanted: Labor*

Markets in Latin America, Baltimore, MD: The Johns Hopkins University Press.

KLEIN, J. P. AND M. L. MOESCHBERGER (1997): *Survival analysis : techniques for censored and truncated data*, New York: Springer.

MALONEY, W. F. AND J. SAAVEDRA-CHANDUVI (2007): "The Informal Sector: What Is It, Why Do We Care, And How Do We Measure It?" in *Informality: Exit and Exclusion*, Washington, D.C.: The World Bank, chap. 1, 21–41.

WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Mass.: The MIT Press.

Chapter 2

Informal Labor Markets and On-the-Job Training: Evidence from Wage Data

2.1 Introduction

This paper explores the role of informal jobs in the formation of human capital among young less-educated workers. An informal job is a job that does not comply with labor regulations. As such, these jobs constitute what is typically known as the informal sector. Traditionally, the informal sector is regarded as the last resort for many workers rationed out of the protected and better paid jobs in the formal sector (e.g. Ozorio de Almeida, Alves, and Graham, 1995), or as the disadvantaged sector in a segmented labor market (e.g. Harris and Todaro, 1970).

This traditional view of the informal sector, however, has been recently challenged by some authors. One example is Maloney (1999). Based on the analysis

of patterns of worker mobility across different sectors of employment, Maloney argues that the existence of an informal labor market in Mexico is not consistent with segmentation in the labor market. Instead, Maloney argues that some workers may be attracted to informal jobs because of their greater flexibility or possibilities for training. Another example is Amaral and Quintin (2006). Following a theoretical approach, Amaral and Quintin show that some of the differences between the formal and informal sectors that are typically interpreted as evidence of barriers of entry into the formal sector, can be an equilibrium outcome in a competitive labor market.

More recently, Arias and Khamis (2008) apply the methods developed in Heckman, Urzua, and Vytlacil (2006) for models with essential heterogeneity to examine the links between earnings performance and the choice of a formal-salaried job, an informal-salaried job, or self-employment.[1] These methods allow Arias and Khamis to account for individuals' observable and unobservable characteristics that influence their decisions to take jobs in one of these sectors. Their results indicate that there is little difference in the earnings of formal-salaried workers and self-employed workers once sorting of workers based on preferences and the returns to their observed and unobserved skills are fully accounted for in the estimation, which is consistent with workers choosing jobs based on their comparative advantages. In contrast, their estimates suggest a clear advantage both for self-employed and formal-salaried

[1]Models with essential heterogeneity are models where responses to interventions are heterogeneous and agents adopt treatments (participate in programs) with at least partial knowledge of their idiosyncratic response.

workers over informal-salaried workers, which is more consistent with the rationing of formal-salaried jobs and with segmentation in the labor market.

The results from household surveys provided by Arias and Maloney (2007) seem to suggest that informal-salaried workers are rationed out of the better paid formal-salaried jobs. Arias and Maloney provide results from household surveys in Argentina, Bolivia, Colombia, and the Dominican Republic that ask employed individuals for the reasons and motivations for taking their current jobs. The results indicate that a substantially higher fraction of informal-salaried workers claimed to have opted for their current job "because they could not find another job," than the fraction of formal-salaried workers (see Table 2.9 of Arias and Maloney, 2007).

Despite this ongoing debate, informal jobs seem to play an important role in the work lives of less-educated workers. Maloney (1999) claims that informal-salaried jobs serve as the main point of entry for young poorly educated workers into paid employment. Following Maloney, this paper focuses on the group of young less-educated workers in Mexico and on their experience in the informal sector. The paper explores the extent to which less-educated workers in the informal sector experience wage growth and how wage growth in the informal sector compares with wage growth in the formal sector. The basic question is whether informal jobs offer wage growth and skill accumulation to less-educated workers, and how it compares with formal jobs.

The empirical analysis uses an employment survey from Mexico, the ENEU.

The panel structure of the survey allows for the construction of measures of wage growth and continuing sector participation. The results indicate that young less-educated workers in the informal sector experience faster wage growth than their peers in the formal sector. Based on existing models of on-the-job training (Becker, 1993; Acemoglu and Pischke, 1999), this result suggests that informal jobs offer valuable general training opportunities to young less-educated workers.

The literature provides little evidence on wage growth in the informal sector or on how it compares with wage growth in the formal sector. There is some evidence on the wage gain (or loss) from informality. For example, Maloney (1999) and Alcaraz, Chiquiar, and Ramos-Francia (2011) provide estimates of the wage change associated with transitions between the formal and informal sectors in Mexico. The results in both studies indicate that informal-to-formal transitions are associated with positive wage changes, while transitions in the opposite direction are associated with negative wage changes. However, none of these two studies provide evidence on wage growth experienced by workers in the informal sector or how it compares to wage growth in the formal sector. The present study contributes to the literature on the informal sector by providing evidence and some suggestive ideas of the mechanisms behind these results.

The study is organized as follows. The following section presents the household survey used in this study and describes the sample and the criteria used to classify employed respondents as formal or informal sector workers. Next,

evidence from wage data is presented in Section 2.3 and the economic interpretations of this evidence are presented in Section 2.4. The last section concludes and discusses future research.

2.2 Data: The ENEU

The empirical analysis is based on data obtained from the Mexican National Survey of Urban Employment, ENEU (its acronym in Spanish). The ENEU is a rotating panel in which households are followed for 12 months, with periodic visits every three months. Consequently, 20% of the sample is replaced every quarter. The empirical analysis in this paper uses data from the third quarter of 1994 to the fourth quarter of 2002; during this period, it is possible to identify 30 different panels, each composed of about 50,000 individuals.

The survey collects information for each individual in the household (e.g. education, sex, position in the family, etc.), and for individuals aged 12 or older, the survey also collects information about their working status and characteristics of their main and secondary jobs. The information on working hours, earnings, benefits, firm size, job position, and industry of occupation refer to the job that the individual held the week prior to the interview. In cases in which the respondent was temporarily absent from work during the week prior to the interview, some information is still collected, but that information does not correspond to the week of reference.[2] As explained in Appendix A.1, this

[2]In some cases, the respondent claims to have a job, but to be absent from work during the

information is used to impute wage data, when this information is missing.

During the period of observation, 47% of all respondents were males. Among the male respondents, about 15% are between the ages of 16 and 20 years (see Figure 2.1), and the average level of education is just below the mandatory level in Mexico, which is grade 9. Figure 2.2 reveals that 34% of males ages 16 to 65 completed primary school (grade 6) and no more; 32% completed middle school (grade 9) and no more; and 16% completed high school but did not go on. About 75% of the male respondents are employees, including salaried and piece-rate workers (see Figure 2.3).

2.2.1 The Sample

The analysis is restricted to males because men and women may have different reasons for opting for a formal or informal job. In particular, one of the most cited reasons by women for choosing an informal-salaried job is the flexibility to work and perform their family duties (Arias and Maloney, 2007). Additionally, the sample only includes salaried and piece-rate workers, not self-employed or employers. However, as Figure 2.4 indicates, 93% of the male respondents between the ages of 16 and 20 are either salaried or piece-rate workers, so the vast majority of the respondents in the age group of interest are employees. Moreover, the sample is restricted to individuals who are salaried or piece-rate workers for the whole time that they are in the survey.

week previous to the interview. This absence from work could be a result of the respondent to be on vacation, on sickness or recovery, or on strike, among other reasons.

To focus on young workers, the sample only includes individuals of ages 16 to 20 inclusive. As explained below, at age 20, transitions between the formal and informal sectors seem to slow down, and so this age is chosen as the upper bound for the sample (see Figure 2.6). On the other hand, age 16 is chosen as the lower bound because of the restrictions imposed by the labor legislation in Mexico.[3] To focus on less-educated individuals, the sample only includes individuals who are not enrolled in school and completed at most the mandatory level of education in Mexico (grade 9).

The sample includes both full-time and part-time workers, although the vast majority of individuals in the sample worked full time. In the sample, 5.56% worked less than 35 hours per week, 68.18% worked between 35 and 48 hours per week, and 26.25% worked more than 48 hours per week.

Finally, the top and bottom 1% of the real hourly earnings are dropped from the sample. The top and bottom percentiles are generated within groups of quarter-year-education, hence there is a different top and bottom percentile for different education levels, on each quarter-year combination.

[3]Article 123-Section III of the Mexican Constitution prohibits the employment of individuals younger than 14 years of age; and for individuals of 14 and 15 years of age it states a maximum of 6 hours of work per day (Constitutional Congress, 1917). Similarly, Article 22 of the Federal Labor Law prohibits employment under 14 years of age, but also the employment of individuals 14 and 15 years of age that have not yet finished the mandatory level of education, which is middle school (Congress, 1970).

2.2.2 Identification of Informal Sector Workers

How is an informal sector worker identified in the sample? In Mexico, labor legislation mandates that all employers register their workers in the Mexican Institute of Social Security, IMSS (its acronym in Spanish).[4] This institution provides a bundle of benefits to registered workers, including: health insurance, day-care services for children, life insurance, disability pensions, work-risk pensions, sports and cultural facilities, retirement pensions, and housing loans (Levy, 2007). Because both the employer and the worker contribute to the IMSS fees, they are motivated not to register or be registered.[5]

Among employees, IMSS is the largest institution providing health insurance. However, there are other institutions providing benefits similar to those of IMSS. One of these institutions is ISSSTE, which provides a bundle of services to state employees, including health insurance. As a result, in Mexico, it is usually said that IMSS or ISSSTE is a benefit associated with one's job. If a worker declares to have health insurance provided by IMSS or ISSSTE, it means that such a worker is a registered worker, and that his or her job abides by the labor regulations. For this reason, the current study uses health insurance provided by IMSS or ISSSTE as the distinguishing feature of formal

[4]Article 123-Section XXIX of the Mexican Constitution states that the Law of Social Security is to the public benefit (Constitutional Congress, 1917). And Article 15-Section I of the Law of Social Security states that every employer must register their employees in the IMSS (Congress, 1995).

[5]The labor law mandates that if the worker earns less than three minimum wages, only the employer pay IMSS fees, but if the worker earns more than three minimum wages, both employer and worker pay these fees.

sector workers, or the lack of it as a distinguishing feature of informal sector workers.

The ENEU's questionnaire asks respondents for the benefits they get from their jobs. The questionnaire provides a list of benefits that the respondent can check. Among these benefits are: IMSS, ISSSTE, paid vacations, Christmas bonus, and private health insurance or other medical services.[6] The respondent can check more than one benefit. For example, a respondent can check both IMSS and "private health insurance or other medical services," which means that the worker is registered in the IMSS, but that also has private health insurance provided by the employer. In this case, the respondent can either use the medical services provided by IMSS or those provided by the private health insurance, or complement the medical services of the private health insurance with those of IMSS, or viceversa.

For the purpose of identifying informal sector workers, this study classifies a respondent as an informal sector worker if the respondent is an employee and neither IMSS nor ISSSTE is checked as an employee benefit. If the respondent checks "private health insurance or other medical services," but neither IMSS nor ISSSTE is checked, the respondent will be classified as an informal sector worker.

It is important to mention that in the questionnaire's option "private health insurance or other medical services," among the "other medical services" are

[6]See question 7d in the ENEU's questionnaire which can be found at `http://www.inegi.org.mx/`.

the medical services for the military and PEMEX employees.[7] As a consequence, the algorithm used in this study to classify workers as informal will classify military and PEMEX employees as informal sector workers, when in fact they are formal sector workers. The proportion of respondents in the sample that does not check IMSS but checks the option of "private health insurance or other medical services," is 2.38%. Hence, the algorithm incorrectly classifies workers as informal in less than 2.38% in the sample.

Figure 2.5 shows the number of workers employed in the formal and the informal sectors by age at the time of the first interview. Notice that, for ages 16 and 17, the majority of less-educated workers are employed in the informal sector, and that for older ages the proportion of workers employed in the formal sector increases. This suggests that, as less-educated workers grow older, they move from the informal into the formal sector.

In fact, Figure 2.6 shows that the likelihood of moving from the informal into the formal sector increases during the first years of the workers' careers. The likelihood of moving in the opposite direction decreases monotonically, suggesting that many workers make the transition from the informal to the formal sector, but as they age, the likelihood that these workers move back to the informal sector decreases. This patten of transitions between these two salaried sectors suggests that young informal-sector workers may expect to eventually move to the formal sector.

[7] PEMEX is the Mexican state-owned petroleum company.

2.3 Evidence from Wage Data

Figure 2.7 presents the kernel density of the log wages of workers in the formal-salaried and informal-salaried sectors in the sample. The kernel densities in the figure are consistent with Arias (2007), who finds that informal-salaried workers have an earnings disadvantage with respect to formal-salaried workers at all points of the pay scale in the case of Argentina and Bolivia. Figure 2.7 suggests that this earnings disadvantage also seems to hold for the case of young less-educated workers in Mexico.

Figure 2.8 presents the evolution of average hourly earnings in the formal and informal sectors during the period of observation. Hourly earnings are in Mexican Pesos of the second half of June 2002. Notice that, during the first periods of observation, hourly earnings fell significantly due to the so-called Tequila crisis. These two series also reflect the greater flexibility in adjusting wages in the informal sector. Both series reach a minimum at the third quarter of 1996, but the loss in hourly earnings in the formal sector is 28%, whereas, in the informal sector, it is 40%. In addition, the growth in hourly earnings between the third quarter of 1996 and the fourth quarter of 2002 is 41% in the formal sector and 56% in the informal sector. Finally, notice that despite the differences in flexibility in adjusting wages in each sector, both series tend to move together, suggesting that they react similarly to changes in economic conditions.

Now, consider individual wages. Table 2.1 presents log-wage regressions for each salaried sector on a set of worker and firm observable characteristics. Most of the estimated coefficients have the expected sign. In both sectors, being a middle-school graduate is much better than only being primary-school graduate, however, the correlation between wages and graduation is stronger in the informal sector, which suggests that for the kind of jobs that less-educated workers access in each sector, skills are more important in the informal sector than in the formal sector. Work experience is positively correlated with wages, as expected.[8] However, given the range of ages, there is not too much curvature in this relationship, hence experience squared is not significant and was not included in the regression. Noticeably, local unemployment has a negative relation with wages in the formal sector, but not in the informal sector.

Notice that industry and firm size are important in explaining wages in both sectors. Figures 2.9 and 2.10 show the distribution of workers in the sample among firms of different sizes and among different industries. Figure 2.9 indicates that formal-salaried workers in the sample are mostly employed in firms with more than 250 employees, whereas, informal-salaried workers are mostly employed in firms with 2 to 5 employees. The fact that some informal-salaried workers are employed in firms with more than 250 employees reveals the well known practice of some firms hiring part of their labor force informally.[9] In such cases, it is typical for the transition from informal to formal to

[8]Experience is computed as $\min\{A - E - 6, A - 16\}$, A =Age, E =Education.

[9]Even though this is a suggestion, it would be very hard to imagine a firm with more than 250 employees and all of them hired informally. In such a case, it would be hard for the

occur within the same firm. Similarly, Figure 2.10 indicates that the majority of young less-educated workers employed formally work in the manufacturing industry. Also, notice that the fraction of workers employed in construction and in services is higher in the informal sector.

Now, consider wage growth. Figure 2.11 shows the kernel density of wage growth in the sample. For both, one-quarter and two-quarter wage growth, wage changes in both sectors are symmetric around 0, but wage growth in the informal sector is more disperse than wage growth in the formal sector. The higher dispersion of wage growth is consistent with the higher flexibility in adjusting wages in the formal sector mentioned above.

Equally important, consider wage growth conditional on worker and firm characteristics. How does individual wage growth in the formal sector compare to wage growth in the informal sector? This relation is explored by estimating the following wage growth equation:

$$\Delta \ln w_{it} = \beta \mathrm{IS}_i + x'_{it}\gamma + \xi_{it} \tag{2.1}$$

where the time index is defined in quarters, $\Delta \ln w_{it} = (\ln w_{it} - \ln w_{it-1})$, IS_i is a dummy for informal-sector participation in two consecutive quarters, and x_{it} is a set of covariates such as those included in the low-wage regressions presented in Table 2.1. The sample used to estimate equation (2.1) only includes workers that are either in the informal sector for two consecutive quarters or are in the formal sector for two consecutive quarters.

employer to stay below the radar of authorities.

The parameter of interest is β, which indicates how wage growth of an individual employed in the informal sector compares to wage growth of an individual employed in the formal sector with similar x_{it} characteristics. Table 2.2 presents the results from estimation of (2.1) for different sets of covariates.

First, consider differences in raw wage growth when x_{it} only includes an intercept. In this case, presented in Column (A) of Table 2.2, the regression results indicate that wage growth in the formal sector is similar to wage growth in the informal sector. Next, consider differences in wage growth when x_{it} includes worker observable characteristics, presented in Column (B). The results indicate that conditioning on education and experience yields the same conclusion. Neither graduation from primary nor secondary school seem to have a significant effect on wage growth. This suggests that educational attainment does not appear to affect wage growth for these workers. Similarly, the effect of experience on wage growth is insignificant.

As Figure 2.8 suggests, it is important to control for different economic conditions over time. To control for these factors, Column (C) presents estimation results including the level of local unemployment and a time trend. Both of theses covariates have significant relationships with wage growth. However, wage growth difference between the two sectors is still insignificant. The estimates indicate that workers in places with higher local unemployment experience lower wage growth. The time trend is intended to capture changes in

economic conditions that affect both sectors. As Figure 2.8 suggests, the economic environment seemed to improve for most of the period of observation. This improvement is reflected in the positive estimate for the time trend.

Finally, Column (D) also controls for characteristics of the firm where the worker is employed by including industry and firm size indicators. The results indicate that industry and firm size are important determinants of wage growth. Furthermore, when controlling for these firm characteristics, the difference in wage growth between the formal and informal sectors becomes positive and significant, indicating that wage growth is faster in the informal sector than in the formal sector.

Industry and firm size indicators are intended to control for differences in firm productivity. If firms of different sizes, or operating in different industries, are systematically different with respect to productivity, then these differences in productivity may lead to differences in wage growth as well as wage levels. One could argue that the larger the firm is, the more productive it is, for example, because larger firms invest more in technology than small firms, and that firms using more technology may require more worker training which will result in higher wage growth. Similarly, one could argue that firms in industries with higher capital to labor ratio could systematically be more productive than firms in other industries.

Recall that formal sector workers are mostly employed in large firms, whereas informal sector workers are mostly employed in small and medium-size firms

(see Figure 2.9). Similarly, the fraction of workers employed in the manufacturing and commerce industries is higher among formal sector workers, whereas, the fraction of workers employed in the construction and services industries is higher among informal sector workers (see Figure 2.10).

Table 2.3 breaks down the estimation of the wage growth equation by industry and by firm size. The numbers in the table give, for different specifications, the estimate of β in equation (2.1), which is the coefficient of the indicator of informal-sector participation in two consecutive quarters, IS_i. The first column indicates that, irrespective of industry, in medium-size firms, informal sector workers experience faster wage growth than formal sector workers, however, the difference in wage growth is not statistically significant for any other firm size. The last two lines indicate that, irrespective of firm size, in the construction and in the services industries, formal sector workers experience faster wage growth than informal sector workers. Breaking down the estimation by industry and firm size, the results indicate that in small and medium-size firms (6 to 10 and 15 to 60 employees), informal sector workers experience faster wage growth than formal sector workers in the construction and service industries.

Notice that none of these specifications control for occupation. It is also possible that less-educated workers in small informal-sector firms are employed in more productive occupations, say mason's apprentice, than less-educated workers employed in small formal-sector firms, say messenger or clerk. Also, recall

that the log-wage equations presented in Table 2.1 suggest that for the kind of jobs that less-educated workers access in each sector, skills seem to be more important in the informal sector. Hence, one might expect wages to grow faster with the acquisition of new skills for less-educated workers in the informal sector than for their peers in the formal sector.

A similar wage growth equation was estimated with $\Delta^2 \ln w_{it} = (\ln w_{it} - \ln w_{it-2})$ and so the indicator IS_i is a dummy for sector participation in three consecutive quarters. Now, the sample used to estimate equation (2.1) only includes workers that are either in the informal sector for three consecutive quarters or are in the formal sector for three consecutive quarters. Table 2.4 presents the estimation results for this specification. Overall, the results and the conclusions are very similar to the one-quarter wage growth: conditional on worker and firm observable characteristics, wages in the informal sector grow faster than wages in the formal sector.

Finally, notice that in all specifications of Tables 2.2 and 2.4, the R^2 is very small, and so a large portion of the variation in wage growth is not explained by the covariates included in the regression. If the omitted variables are systematically correlated with informal or formal sector participation, then the indicator for continuous informal sector participation will pick up these correlation.

Similarly, there is no explicit treatment of unobserved heterogeneity or selection, and so the estimates on the indicator of informal sector participation

could be biased. However, if unobserved heterogeneity, or "ability," has a similar effect on wages in two or three consecutive quarters, then its effect should cancel out when looking at wage growth. Notice that educational achievement does not have a significant effect on wage growth, and so it seems unlikely that unobserved heterogeneity would have a crucial role, given the strong correlation between education and unobserved heterogeneity (or ability). With respect to worker selection, one can easily argue that those workers continuously employed in the informal sector are negatively selected, and so the estimate of the coefficient of IS_i may be downward biased. In this case, we can consider it as a lower bound for the true β.

2.4 Economic Interpretations of Evidence

This section argues that this evidence is consistent with general human capital investment on-the-job.[10] First, consider the model of general on-the-job training in a competitive labor market provided by Becker (1993). In such a labor market, wages paid by a firm are determined by the productivity in other firms. Productivity increases with general training equally in the firm providing it as well as in other firms. Consequently, firms cannot capture any of the returns from the investment in general training because the worker can move freely to another firm once training is finished. As a result, workers capture all the

[10]General training increases a worker's productivity at any firm. Contrary to firm-specific training, which increases productivity more in firms providing it. See Becker (1993) chapter III.

returns from that investment and bear the cost of general training.

Acemoglu and Pischke (1999) show that, if frictions in the labor market result in a *compressed wage structure*, firms find it profitable to invest in training, even when training involves general skills. In a compressed wage structure, productivity in the current firm increases more with training than in other firms. Hence, firms' profits increase with training, as a consequence firms are willing to sponsor general training.

Acemoglu and Pischke (1999) also provide examples of mechanisms that produce a compressed wage structure, inducing firms to sponsor general training. Some of these mechanisms include search frictions that generate job search costs, asymmetric information about the worker's ability, complementarity between firm-specific skills and general skills, and labor market institutions, such as minimum wages and unions. Equally important, Acemoglu and Pischke show that increasing wage compression leads to more firm-sponsored training.

The informal labor market is likely to be more competitive than the formal labor market. This feature of the labor market was exploited by Zenou (2008). Zenou develops a model of the informal and formal sectors in which the formal labor market is characterized by search frictions, while the informal labor market is competitive. Equally important are frictions generated by labor institutions. One of the most cited causes of large informal sectors is the existence of rigidities in the labor market due to excessive regulation (see Schneider and

Enste, 2000). This link between labor regulations and the existence of informal sectors has been studied and documented. For example, Rauch (1991) develops a model in which the size of the informal sector is directly related to the degree of labor regulation.[11] Bosch, Goni, and Maloney (2007) find that the main driving force behind the increase in informality in Brazil during the 1990s was the reduction of formal sector hirings mainly explained by changes in labor market legislation. More recently, Albrecht, Navarro, and Vroman (2009) built an equilibrium search and matching model to study the effects of changes of severance and payroll taxes; their simulations suggest that increases in both severance and payroll taxes shift employment from the formal to the informal sector.

Wage compression in the formal labor market due to frictions implies that firms reap some of the returns from training and pay at least part of the cost of training. The informal labor market is more competitive, hence workers reap the returns from training and bear the cost of training. As a consequence, even with the same amount of investment on training in both sectors, wage growth should be faster in the informal sector than in the formal sector.

It is also possible that informal sector workers invest more in human capital than formal sector workers. Figures 2.7 and 2.8 show that, on average, wages in the informal sector are lower than in the formal sector. The difference in intercepts in the log-wage equations in Table 2.1 suggests that there is still

[11] Labor regulation in Rauch (1991) is implemented as a minimum wage. Acemoglu and Pischke (1999) argue that this is one of the mechanisms producing a compressed wage structure.

a gap after controlling for worker and firm observable characteristics. Finally, Figure 2.6 suggests that, during the first years of the workers' careers, workers are more likely to move from the informal to the formal sector as they age. If wages and productivity are lower in the informal sector and informal sector workers expect to move to the formal sector eventually, investment in human capital may be greater in the informal sector than in the formal sector. That is, if workers face a lower price for their skills in the informal sector in the present, and expect a higher price for their skills when they move to the formal sector in the future, then the opportunity cost of human capital investment is lower in the informal sector, which will induce informal sector workers to invest more in human capital. Formal sector workers, on the other hand, do not face this lower opportunity cost.

Faster wage growth in the informal sector could also arise in a model of on-the-job training with different levels of specificity of training in the formal and informal sectors. In Becker (1993), when firms provide firm-specific on-the-job training, firms bear the cost of training because if the worker moves to another firm, all productivity gains from training will be lost. If most of the training in the formal sector involves firm-specific human capital, whereas most of the training in the informal sector involves general human capital, wages in the informal sector will exhibit greater growth than wages in the formal sector.

2.5 Final Remarks

The traditional view of the informal sector assumes that jobs in this sector offer little beyond a make-shift or temporary job for workers that are waiting for a "better" formal sector job. However, this study shows that informal jobs are not dead end jobs, and that these jobs appear to offer wage growth similar to formal sector jobs for young less-educated workers entering the labor market.

The present study provides an analysis of the informal and formal sectors using data from 1994 to 2002 from the Mexican National Survey of Urban Employment, ENEU. The analysis revealed that less-educated workers start their careers in the informal sector, and move to the formal sector as they grow older. More important, it is found that for young less-educated workers, wages in the informal sector grow faster than wages in the formal sector, conditional on worker and firm observable characteristics.

On the assumption that the labor market in the informal sector is more competitive than its counterpart in the formal sector, models of on-the-job training in competitive and in non-competitive labor markets predict that formal sector employers sponsor at least part of the training costs, while informal sector employers pass these costs onto the workers. The evidence from wage growth data presented in this paper is consistent with these theories of human capital accumulation, which, in turn, supports the possibility that young less-educated workers accumulate skills while employed in the informal sector.

Informal-salaried workers may even invest more in human capital than formal-salaried workers. This is because informal-salaried workers have lower wages and expect to eventually move to a formal-salaried job. Another mechanism consistent with the evidence on wage data is based on systematic differences in the specificity of training between the informal and the formal sectors. Even though all these mechanisms lead to the same conclusion in terms of investment in human capital in the informal sector, they may have different implications for the design of labor market policies, and so ideally one could distinguish between them. That exercise would require more than just analyzing wage data.

If informal-salaried jobs do indeed provide provide training to young less-educated workers, those who start in the informal sector and move to the formal sector later on in their careers, will have a career path different from the career path of less-educated workers who start in the formal sector. These differences could be used to distinguish between the proposed mechanisms discussed above. However, to study these differences, it is necessary to have access to a longer panel than the one used in this paper, which only follows individuals during 12 months.

Equally important, if informal jobs provide training opportunities to young less-educated workers, then it is possible that these workers opt for an informal job instead of queuing longer for a formal job, in order to accumulate skills. Given the arguments provided before, it could also be possible that training

costs could help in closing the gap between earnings in the formal and informal sectors, which seem to persist after controlling for observable characteristics.

It is true that this study only explores one possible role of the informal sector in the careers of less-educated workers, human capital accumulation. However, the informal sector may have other roles. For example, the informal sector could play the role of a screening device. That is, suppose that when less-educated workers enter the labor market their abilities are unknown, and so, to minimize firing costs, formal sector firms refuse to hire them. If the informal sector offers job opportunities to young less-educated workers, and their ability is revealed while working there, then formal sector firms could use the worker's trajectory in the informal sector to learn the worker's ability and hire from the pool of informal sector workers whose ability has been revealed.

Chapter 3 considers two roles of the informal sector: human capital accumulation and screening of workers' abilities. Based on the implications of a search and matching model, and on the estimation of the hazard function from informal to formal sectors, the author concludes that the main role of informal jobs is to serve as a screening mechanism that solves an information problem about workers' abilities. Although this result does not rule out the possibility of workers accumulating skills in the informal sector it has important implications for the design of labor market policies directed to the informal labor market.

2.6 Bibliography

ACEMOGLU, D., AND J.-S. PISCHKE (1999): "The Structure of Wages and Investment in General Training," *The Journal of Political Economy*, 107(3), 539–572.

ALBRECHT, J., L. NAVARRO, AND S. VROMAN (2009): "The Effects of Labour Market Policies in an Economy with an Informal Sector," *The Economic Journal*, 119(539), 1105–1129.

ALCARAZ, C., D. CHIQUIAR, AND M. RAMOS-FRANCIA (2011): "Wage differentials in Mexico's urban labor market," *Economics Bulletin*, 31(3), 2500–2508.

AMARAL, P. S., AND E. QUINTIN (2006): "A competitive model of the informal sector," *Journal of Monetary Economics*, 53, 1541–1553.

ARIAS, O. (2007): "Informality, Earnings, and Welfare," in *Informality: Exit and Exclusion*, chap. 3, pp. 79–100. The World Bank, Washington, D.C.

ARIAS, O., AND M. KHAMIS (2008): "Comparative Advantage, Segmentation and Informal Earnings: A Marginal Treatment Effects Approach," Discussion Paper IZA DP No. 3916, The Institute for the Study of Labor (IZA).

ARIAS, O., AND W. F. MALONEY (2007): "The *Razón de Ser* of the Informal Worker," in *Informality: Exit and Exclusion*, chap. 2, pp. 43–78. The World Bank, Washington, D.C.

BECKER, G. S. (1993): *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. Chicago: The University of Chicago Press (for NBER), 3d edn.

BOSCH, M., E. GONI, AND W. F. MALONEY (2007): "The Determinants of Rising Informality in Brazil: Evidence from Gross Worker Flows," Working Paper 4375, The World Bank.

CONGRESS (1970): *Federal Labor Law*. Mexico, D.F.

CONGRESS (1995): *Law of Social Security*. Mexico, D.F.

CONSTITUTIONAL CONGRESS (1917): *Political Constitution of the United Mexican States*. Constitutional Congress, Mexico, D.F.

HARRIS, J. R., AND M. P. TODARO (1970): "Migration, Unemployment and Development: A Two-Sector Analysis," *The American Economic Review*, 60(1), 126–142.

HECKMAN, J. J., S. URZUA, AND E. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

LEVY, S. (2007): "Can Social Programs Reduce Productivity and Growth? A Hypothesis for Mexico," Mimeo.

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

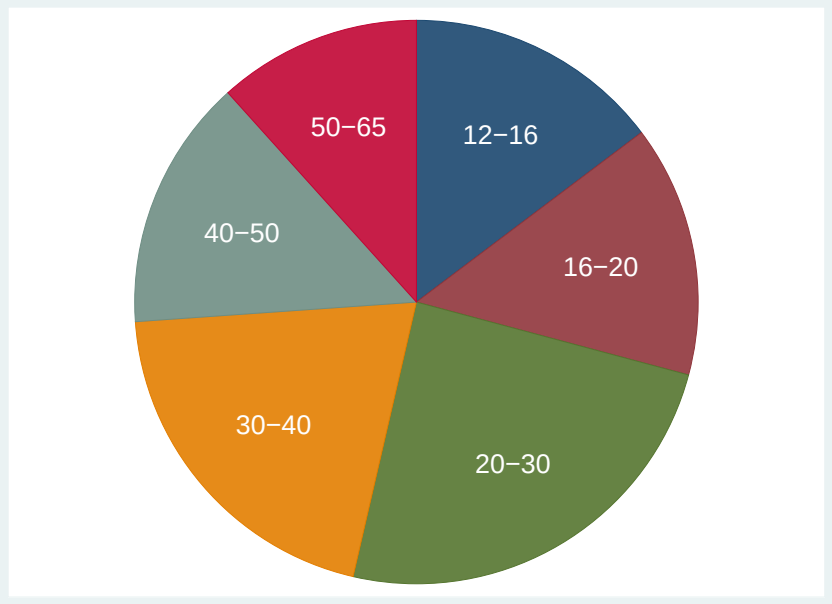
OZORIO DE ALMEIDA, A. L., L. ALVES, AND S. E. M. GRAHAM (1995): "Poverty, Deregulation, and Employment in the Informal Sector of Mexico," ESP Discussion Paper 54, Education and Social Policy Department, World Bank, Washington, DC.

RAUCH, J. E. (1991): "Modeling the informal sector formally," *Journal of Development Economics*, 35(1), 33–47.

SCHNEIDER, F., AND D. H. ENSTE (2000): "Shadow Economies: Size, Causes, and Consequences," *Journal of Economic Literature*, 38(1), 77–114.

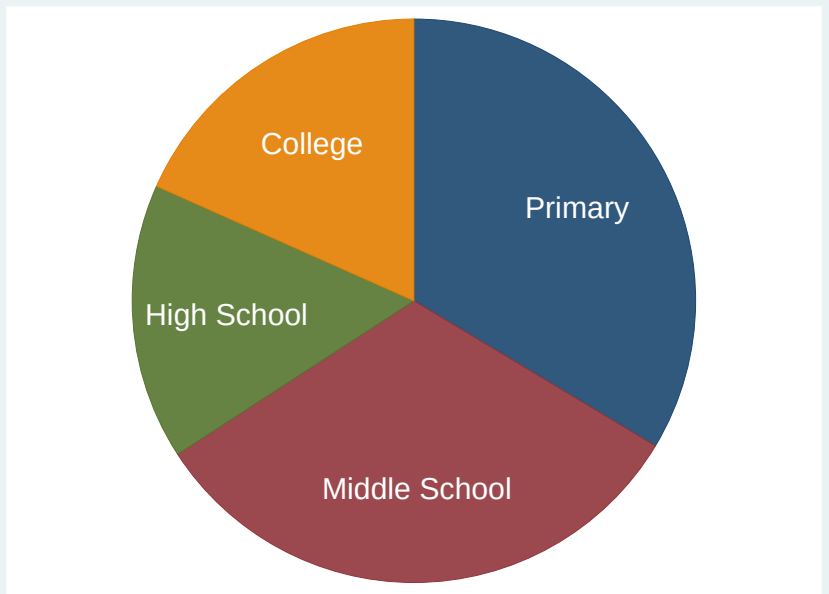
ZENOU, Y. (2008): "Job search and mobility in developing countries. Theory and policy implications," *Journal of Development Economics*, 86(2), 336–355.

Figure 2.1: Age Distribution in the ENEU



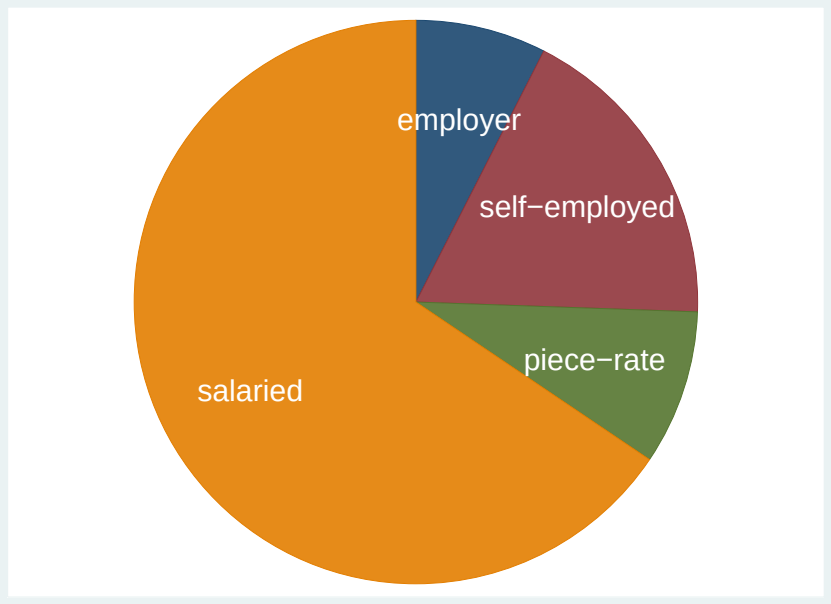
Source: ENEU 3:1994 - 4:2002. Includes only males ages 12 to 65

Figure 2.2: Education Distribution in the ENEU



Source: ENEU 3:1994 - 4:2002. Includes only males ages 12 to 65

Figure 2.3: Job Position Distribution in the ENEU

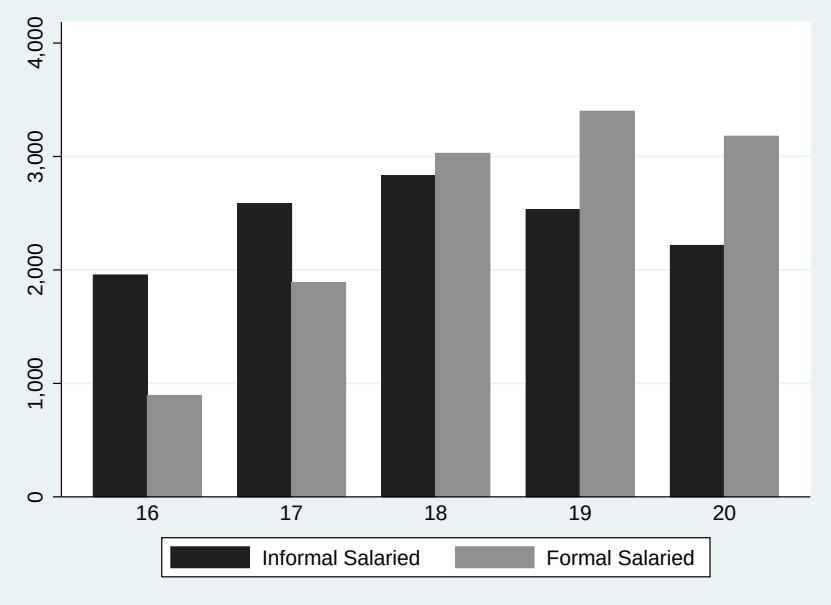


Source: ENEU 3:1994 - 4:2002. Includes only males ages 12 to 65

Figure 2.4: Job Position Distribution Ages 16 - 20

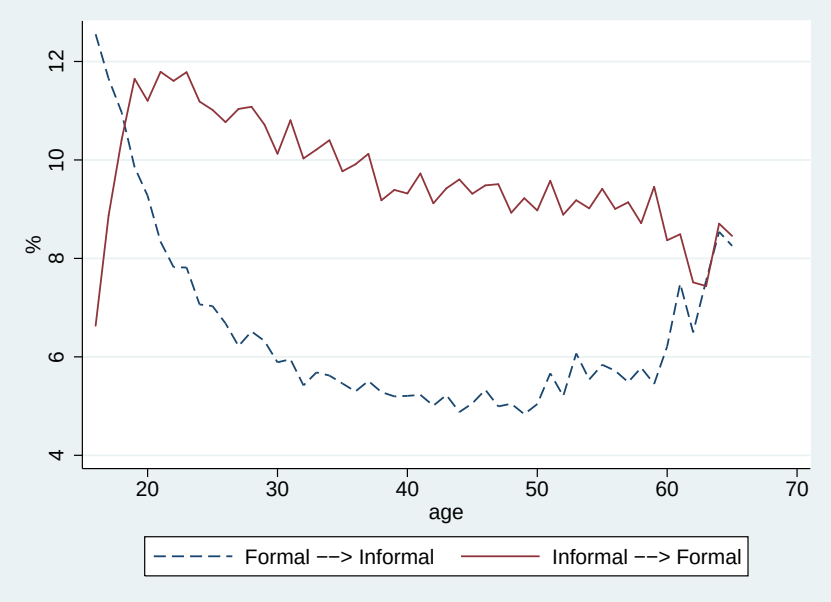
Source: ENEU 3:1994 - 4:2002. Includes only males ages 16 to 20

Figure 2.5: Formal Salaried and Informal Salaried Workers by Age



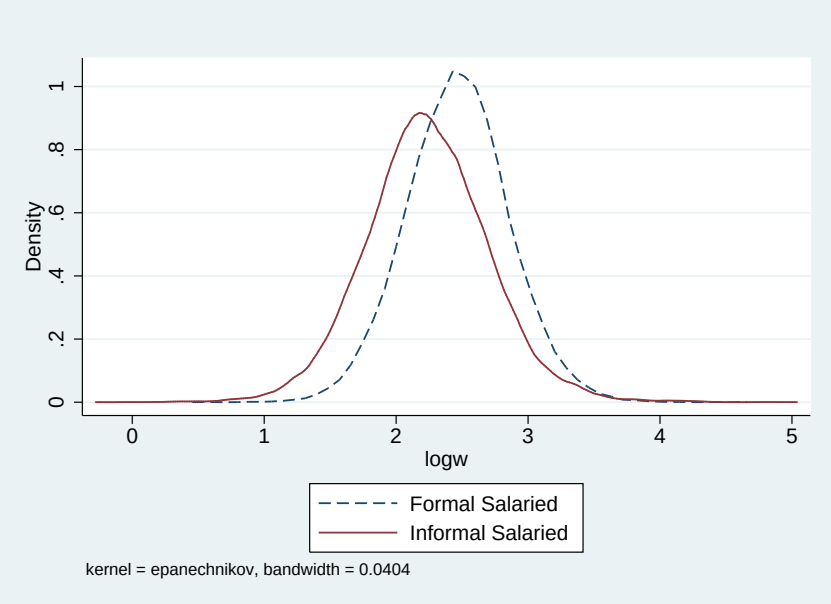
Source: ENEU 3:1994-4:2002. Males only, with 0 to 9 years of education and with no changes in the level of education.

Figure 2.6: Worker Transitions by Age as a Fraction of Initial Sector



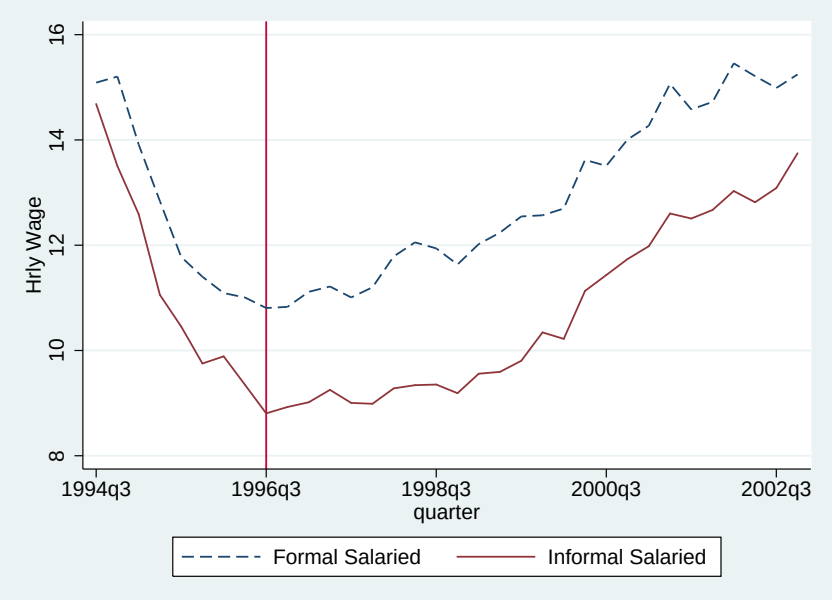
Source: Author's calculations using ENEU. FS = Formal Salaried, IS = Informal Salaried. Ages 16 to 65.

Figure 2.7: Kernel Density of Log-Wages in the Sample



Source: Author's calculations using ENEU. Includes only males ages 16 to 20 with education less or equal to 9 years and with no changes in education level.

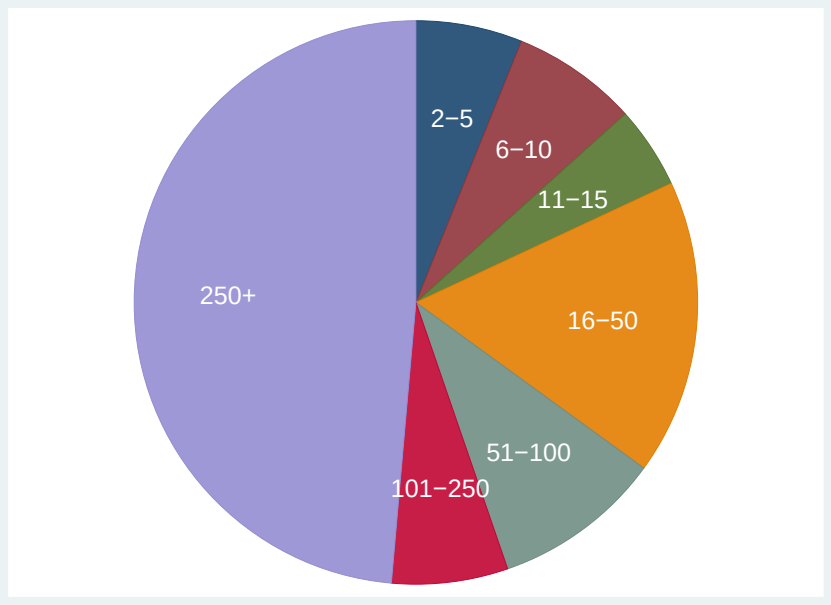
Figure 2.8: Average Wage over Time by Sector in the Sample



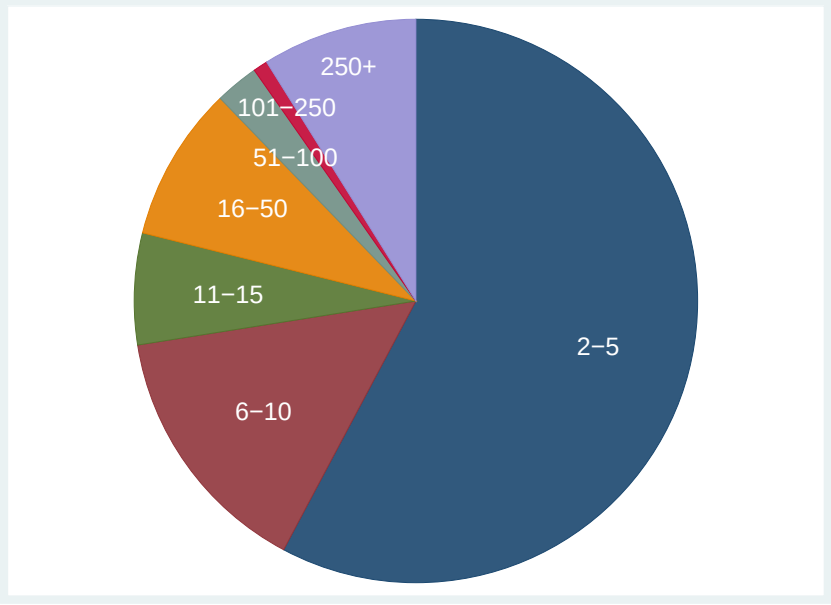
Source: Author's calculations using ENEU. Includes only males ages 16 to 20 with education less or equal to 9 years and with no changes in education level. Hourly wage in Mexican pesos as in the second-half of June 2002.

Figure 2.9: Firm Size Distribution in the Sample

(a) Formal Sector



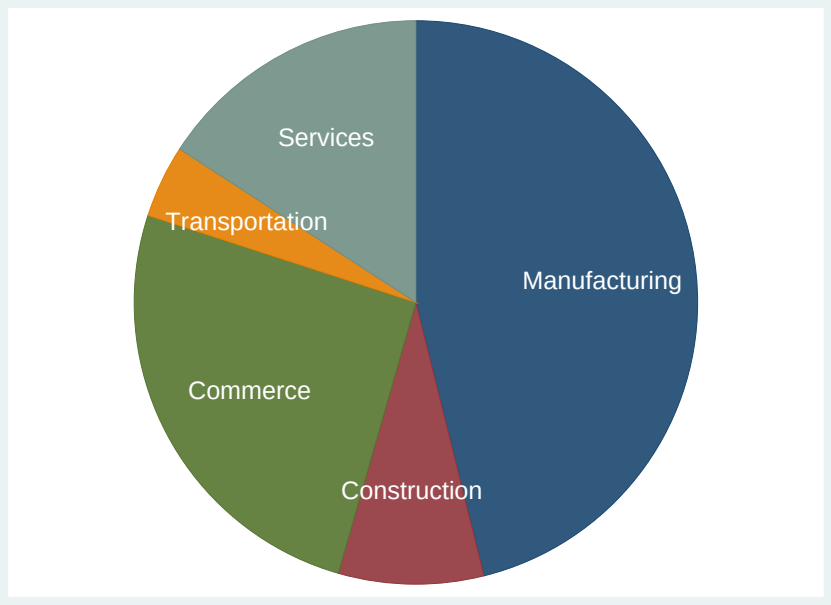
(b) Informal Sector



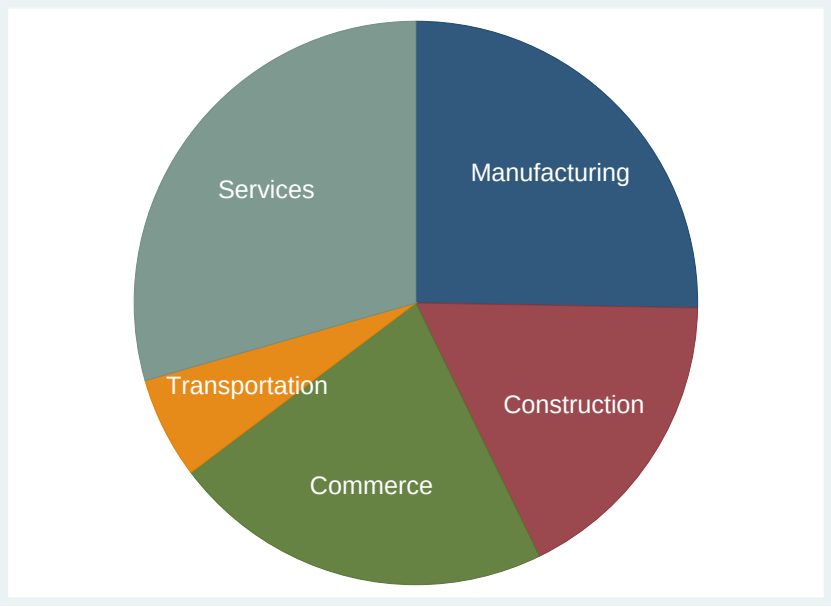
Source: ENEU 3:1994-4:2002. Includes only males ages 16 to 20 with education less or equal to 9 years and with no changes in education level.

Figure 2.10: Industry Distribution in the Sample

(a) Formal Sector



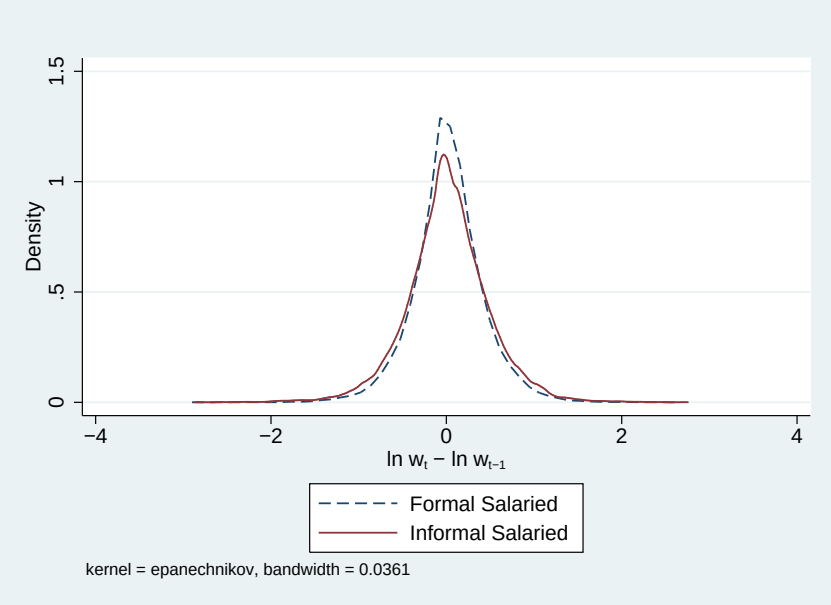
(b) Informal Sector



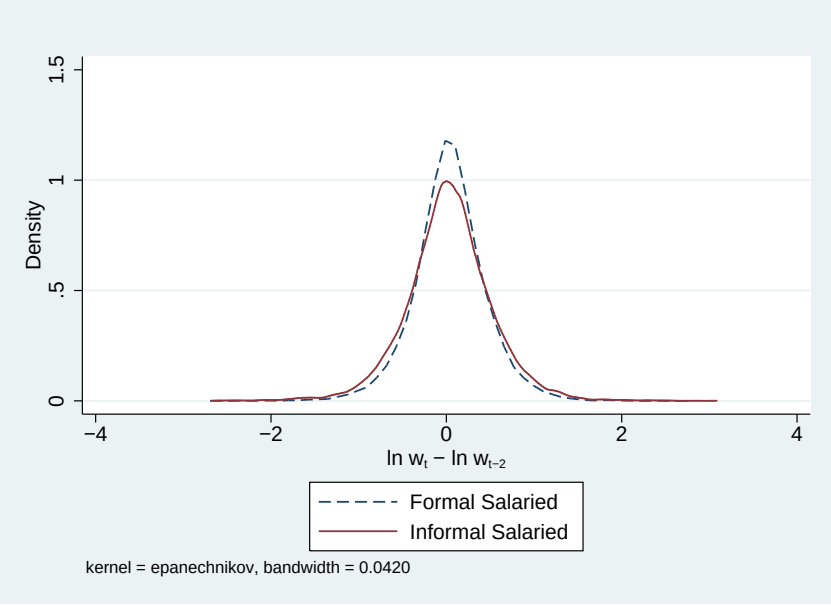
Source: ENEU 3:1994-4:2002. Includes only males ages 16 to 20 with education less or equal to 9 years and with no changes in education level.

Figure 2.11: Kernel Density of Wage Growth in the Sample

(a) One-Quarter



(b) Two-Quarters



Source: Author's calculations using ENEU. Includes only males ages 16 to 20 with education less or equal to 9 years and with no changes in education level.

Table 2.1: Log-Wage Regressions by Sector

	Informal Sector	Formal Sector
Primary School Grad.	0.0759	0.0189
	(0.0072)	(0.0082)
Middle School Grad.	0.0375	0.0277
	(0.0050)	(0.0039)
Experience	0.0471	0.0304
	(0.0017)	(0.0015)
Local Unemployment	-0.1861	-1.2361
	(0.1375)	(0.1161)
Time trend	0.0085	0.0077
	(0.0003)	(0.0003)
INDUSTRY		
Construction	0.0783	0.0503
	(0.0073)	(0.0082)
Commerce	-0.0481	-0.1151
	(0.0063)	(0.0043)
Services	-0.0073	-0.0647
	(0.0065)	(0.0057)
FIRM SIZE		
6-10	0.0873	0.0317
	(0.0068)	(0.0097)
11-15	0.1315	0.0543
	(0.0098)	(0.0108)
16-50	0.1150	0.0468
	(0.0084)	(0.0084)
51-100	0.1116	0.0754
	(0.0150)	(0.0092)
101-250	0.1092	0.0757
	(0.0240)	(0.0100)
250+	0.2019	0.1352
	(0.0086)	(0.0077)
constant	1.8550	2.1988
	(0.0122)	(0.0132)
Number of obs.	38904	42880
R^2	0.0764	0.0974

NOTES: Primary School Grad.$= \mathbb{I}\{E \geq 6\}$, and Middle School Grad.$= \mathbb{I}\{E \geq 9\}$, where E is years of education. The omitted industry is Manufacturing, and the Transportation industry was included in the Commerce industry. The omitted firm size is 2-5 employees. The sample includes males ages 16 to 20 years of age not enrolled in school with 9 or less years of education. Standard errors of estimates are in parenthesis.

Table 2.2: One-Quarter Wage Growth Regressions

	(A)	(B)	(C)	(D)
IS	-0.0063	-0.0063	-0.0045	0.0106
	(0.0040)	(0.0042)	(0.0042)	(0.0058)
Primary School Grad.		0.0008	0.0009	0.0026
		(0.0073)	(0.0073)	(0.0073)
Middle School Grad.		0.0027	0.0014	0.0017
		(0.0043)	(0.0043)	(0.0043)
Experience		-0.0012	-0.0006	-0.0005
		(0.0016)	(0.0016)	(0.0016)
Local Unemployment			-0.4233	-0.4059
			(0.1225)	(0.1230)
Time trend			0.0020	0.0020
			(0.0003)	(0.0003)
INDUSTRY				
Construction				0.0306
				(0.0074)
Commerce				-0.0036
				(0.0050)
Services				0.0057
				(0.0057)
FIRM SIZE				
6-10				0.0326
				(0.0074)
11-15				0.0340
				(0.0099)
16-50				0.0302
				(0.0077)
51-100				0.0368
				(0.0100)
101-250				0.0314
				(0.0116)
250+				0.0359
				(0.0069)
constant	0.0190	0.0200	-0.0057	-0.0415
	(0.0027)	(0.0084)	(0.0114)	(0.0130)
Number of obs.	44,754	44,754	44,754	44,754
R^2	0.0001	0.0001	0.0031	0.0043

NOTES: IS_i is and indicator for continuous participation in the informal sector. $\text{IS}_i = 1$ if individual participated two consecutive quarters in the informal sector, and $\text{IS}_i = 0$ if the individual participated two consecutive quarters in the formal sector. Primary School Grad.$= \mathbb{I}\{E \geq 6\}$, and Middle School Grad.$= \mathbb{I}\{E \geq 9\}$, where E is years of education. The omitted industry is Manufacturing, and the Transportation industry was included in the Commerce industry. The omitted firm size is 2-5 employees. The sample includes males ages 16 to 20 years of age not enrolled in school with 9 or less years of education. Standard errors of estimates are in parenthesis.

Table 2.3: Wage Growth Regressions: Coefficient of Informal Sector Participation for Two Consecutive Quarters

		Industry			
Firm Size	All Industries	Manufacturing	Construction	Services	Commerce
2-5	0.015	-0.012	-0.040	0.012	0.032
	(0.015)	(0.030)	(0.059)	(0.031)	(0.023)
6-10	0.004	-0.007	0.024	0.063**	-0.033
	(0.014)	(0.028)	(0.052)	(0.031)	(0.023)
11-15	0.009	0.010	-0.036	0.000	0.011
	(0.018)	(0.029)	(0.063)	(0.054)	(0.031)
16-50	0.028**	0.030	0.076*	0.057	-0.013
	(0.013)	(0.019)	(0.041)	(0.038)	(0.024)
51-100	0.012	0.026	0.044	0.004	-0.026
	(0.021)	(0.031)	(0.059)	(0.067)	(0.040)
101-250	0.008	0.049	0.005	-0.079	-0.016
	(0.036)	(0.049)	(0.108)	(0.118)	(0.084)
250+	0.008	0.018	-0.085	0.009	-0.033
	(0.010)	(0.029)	(0.068)	(0.017)	(0.029)
All Firm Sizes		0.001	-0.036**	-0.015*	0.000
		(0.007)	(0.016)	(0.008)	(0.011)

NOTES: All regressions include the same covariates as the regression in column (D) of Table 2.2. Standard errors of estimates are in parenthesis.

** Significant at 5%, * Significant at 10%

Table 2.4: Two-Quarters Wage Growth Regressions

	(A)	(B)	(C)	(D)
IS	-0.0050	-0.0049	-0.0012	0.0168
	(0.0057)	(0.0060)	(0.0060)	(0.0087)
Primary School Grad.		-0.0081	-0.0083	-0.0065
		(0.0107)	(0.0107)	(0.0107)
Middle School Grad.		0.0110	0.0088	0.0092
		(0.0062)	(0.0061)	(0.0062)
Experience		-0.0030	-0.0021	-0.0020
		(0.0024)	(0.0024)	(0.0024)
Local Unemployment			-1.0516	-1.0283
			(0.1734)	(0.1741)
Time trend			0.0037	0.0037
			(0.0004)	(0.0004)
INDUSTRY				
Construction				0.0255
				(0.0108)
Commerce				-0.0165
				(0.0071)
Services				0.0019
				(0.0082)
FIRM SIZE				
6-10				0.0289
				(0.0109)
11-15				0.0010
				(0.0147)
16-50				0.0369
				(0.0114)
51-100				0.0365
				(0.0146)
101-250				0.0249
				(0.0166)
250+				0.0364
				(0.0101)
constant	0.0316	0.0411	0.0035	-0.0276
	(0.0039)	(0.0126)	(0.0167)	(0.0191)
Number of obs.	22,839	22,839	22,839	22,839
R^2	0.0000	0.0002	0.0121	0.0136

NOTES: IS_i is and indicator for continuous participation in the informal sector. $\text{IS}_i = 1$ if individual participated three consecutive quarters in the informal sector, and $\text{IS}_i = 0$ if the individual participated three consecutive quarters in the formal sector. Primary School Grad.$= \mathbb{I}\{E \geq 6\}$, and Middle School Grad.$= \mathbb{I}\{E \geq 9\}$, where E is years of education. The omitted industry is Manufacturing, and the Transportation industry was included in the Commerce industry. The omitted firm size is 2-5 employees. The sample includes males ages 16 to 20 years of age not enrolled in school with 9 or less years of education. Standard errors of estimates are in parenthesis.

Chapter 3

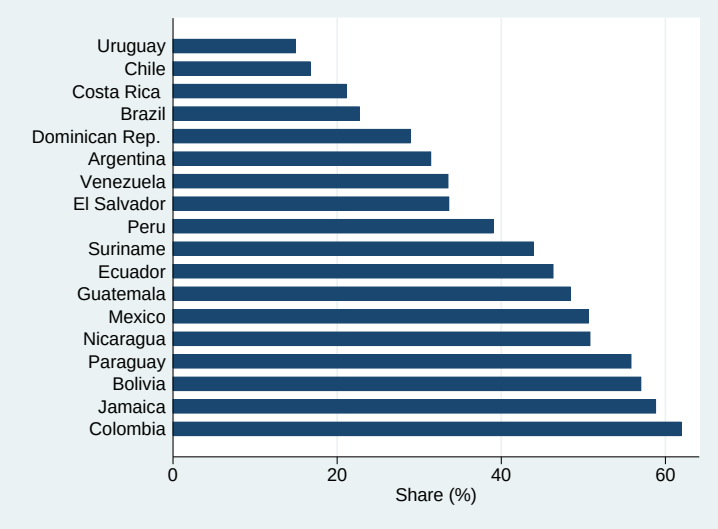
The Role of the Informal Sector in the Early Careers of Less-Educated Workers

3.1 Introduction

The informal sector is an important feature of labor markets in developing countries. This sector, composed of all jobs not complying with labor regulations, occupies a significant portion of these countries' labor markets. In Latin America and the Caribbean, the fraction of workers employed in the informal sector ranges from 15% to 62% (see Figure 3.1). Jobs in this sector employ the majority of young unskilled workers usually paying very low wages, not to mention the lack of health and employment insurance enjoyed by workers holding formal sector jobs.

The presence of large informal sectors has typically been a concern for researchers and policymakers. Some are concerned that the informal sector could

Figure 3.1: Share of Salaried Workers in Informal Jobs in Latin America and the Caribbean



Source: Socio-Economic Database for Latin America and the Caribbean (CEDLAS and The World Bank). Data obtained in the fall of 2010. Males and females ages 25-64 in urban areas. Varying years. A worker is considered informal if (s)he does not have the right to a pension when retired.

be the disadvantaged sector in a segmented labor market market (Magnac, 1991; Maloney, 1999; Amaral and Quintin, 2006; Arias and Khamis, 2008). Others are concerned that the informal sector might adversely affect productivity and growth (Loayza, 1996; Schneider and Enste, 2000; Farrell, 2004; Levy, 2007; Fajnzylber, 2007). Whether these concerns are supported by the evidence is still unresolved. However, they have induced policymakers to introduce tighter regulations to reduce or control the size of the informal sector.

Before attempting to restrict the informal sector, it is important to investigate the potential benefits that workers obtain during informal sector employment. Previous studies have found that less-educated workers start their working careers in salaried jobs in the informal sector and move into formal

jobs as they grow older (Maloney, 1999; Arias and Maloney, 2007). We would like to know if informal sector jobs provide some value above and beyond makeshift low-paying work while people wait to find a "good" formal sector job: do these jobs also provide skills or help screen workers to facilitate a transition to higher paying formal sector jobs? If rules designed to reduce the informal sector are implemented, would we lose some valuable worker training or screening? If so, restrictions on informal sector employment should be accompanied by policies that replace the productive functions of these jobs.

We investigate two potential roles that informal sector jobs could play in the early stages of a worker's career. First, these jobs may provide the opportunity to accumulate skills, making workers more productive and more attractive to formal sector employers. While more-educated workers tend to access greater training opportunities in formal sector employment, less-educated workers may turn to the informal sector to gain work skills.[1] Second, informal sector jobs may serve as a screening device that enables employers to learn a worker's ability. The lack of compliance with labor regulations, especially firing costs and severance payments, suggests that informal sector employers may be more prone to hire young unskilled workers entering the labor market than are formal sector employers. Hence, an informal sector worker who reveals that he is productive may increase his likelihood of finding a formal sector job.

The role of the informal sector as a provider of training opportunities was

[1]The evidence presented by Barron, Berger, and Black (1997) indicates that more educated workers in the U.S. have greater access to on-the-job training (see Table 4.2).

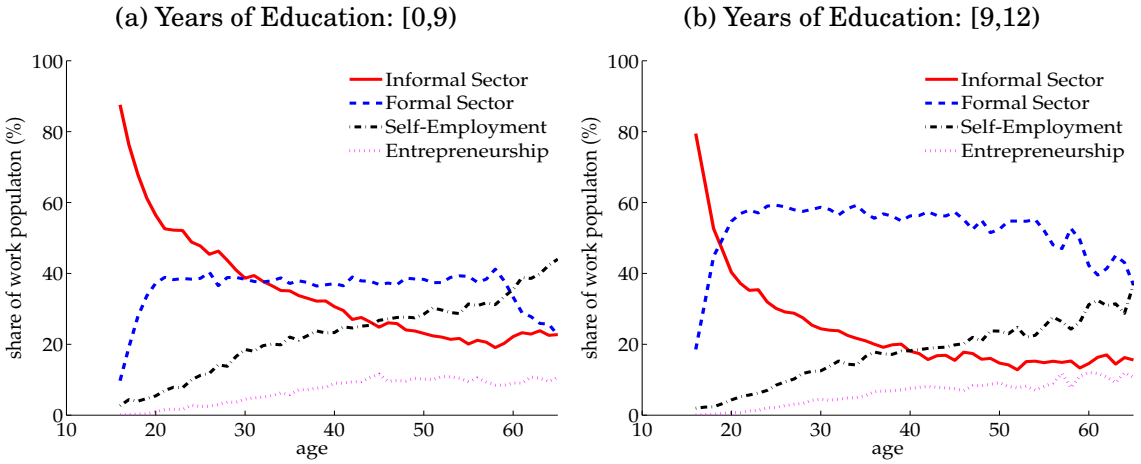
first suggested by Hemmer and Mannel (1989) and has been advocated by Maloney (1999) and Arias and Maloney (2007). The role of the informal sector as a screening device is rarely discussed. One exception is Arias and Maloney (2007) who argue that labor regulations and information asymmetries "impede young workers' entry into the formal sector."[2] The study presented here contributes to this literature by providing an analytical framework and empirical evidence about these roles of the informal sector.

To determine the relative importance of the training or screening roles of the informal sector, we develop a two-sector matching model to study worker movements from the informal to the formal sector. The model is designed to better understand the labor market dynamics in Mexico, a country with a significant informal labor market. In Mexico, the informal sector is a port of entry to the labor market for less-educated workers. These workers are concentrated in the informal sector in the early stages of their working careers, moving to the formal sector as they age (see Figure 3.2). Figure 3.3 shows that the probability of moving from the informal to the formal sector increases during the early stages of workers' careers.

The empirical analysis is based on the analytical implications for hazard rates from the informal to the formal sectors derived from the model. It is shown that hazard rates from informal to formal sectors as a function of tenure

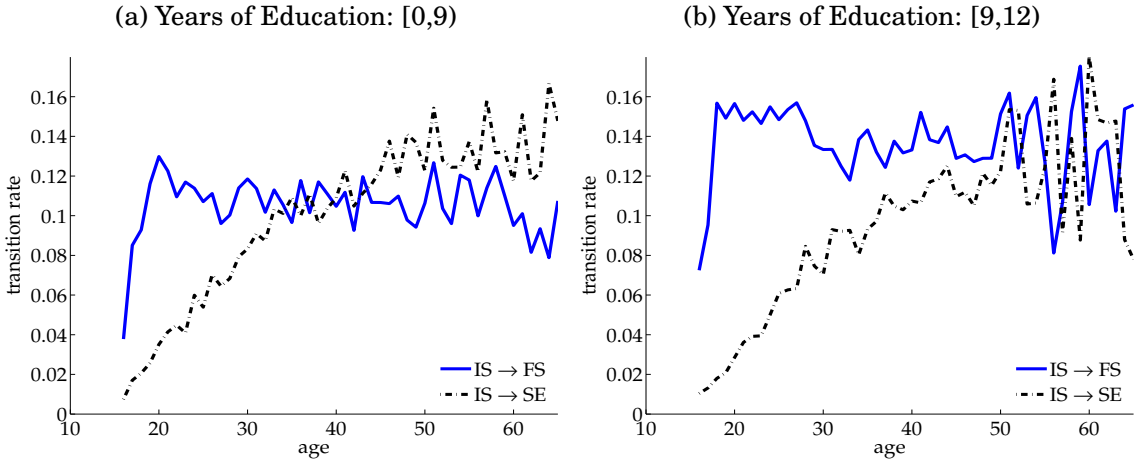
[2]Bosch (2006) and Bosch, Goni, and Maloney (2007) present evidence that labor regulations affect the patterns of job creation in the formal sector in economies with large informal sectors. Some argue that these regulations disproportionately affect the youth (World Bank, 2007, chap. 4).

Figure 3.2: Distribution of Workers by Employment Sector in Mexico



Source: Author's calculations using ENOE I:2005 - IV:2010. A worker is considered informal if he is an employee not enrolled in government health care program. Males not attending school.

Figure 3.3: Transitions Out of the Informal Sector in Mexico



Source: Author's calculations using ENOE I:2005 - IV:2010. Number of transitions relative to the size of the informal sector. A worker is considered informal if he is an employee not enrolled in government health care program. Males not attending school. IS = Informal Sector, FS = Formal Sector, SE = Self-Employed.

differ depending on the role of the informal sector: human capital accumulation or screening. On the one hand, if workers accumulate human capital while working in the informal sector, the likelihood of moving into the formal sector increases with informal sector tenure. On the other hand, if workers' productivities are screened while working in the informal sector, those discovered as highly productive move faster to the formal sector, leaving behind those with low productivity who have difficulties to access formal sector jobs. Thus, the likelihood of moving into the formal sector may initially increase, but it eventually decreases with informal sector tenure.

Using an employment survey from Mexico to obtain measures of duration of employment in the informal sector, we estimate the hazard functions and test the two hypotheses. The estimated hazard is consistent with the implications of the screening model, which indicates that informal sector jobs have an important role by solving the information problem about the abilities of young less-educated workers that are new to the labor market.

Our results give us the means to infer the parameters governing the screening process in one stream of the *Bécate* training program for the unemployed in Mexico, which is targeted at less-educated youth.[3] One of the streams of *Bécate* is a mixture of skill formation and worker placement. In this stream, training takes place at the workplace, and the hosting firm must have empty

[3] *Bécate* was launched in 1984 and was designed to assist individuals with less than 9 years of education between the ages of 16 and 30. Currently, the program has more streams to assist a broader set of workers and needs. Delajara, Freije, and Soloaga (2006) provides a comprehensive evaluation of the program.

vacancies that need to be filled. The training program lasts for one to three months. At the end of the training program, the firm is committed to hire at least 70% of the participants.[4] Given this short amount of time, it seems likely that the program works more as a screening device than a source of significant skill formation.

Based on the estimated hazard, we can deduce the rate at which an employer learns about a worker's ability. For workers with less than 12 years of education, the estimates indicate that an employer learns about a worker's ability at a rate of 14% per month. Consequently, if an employer commits to hire 70% of the program participants, a one or two month program requires the employer to take a gamble on a considerable portion of the program participants, since the employer must bear the firing costs of terminating any unsuitable workers. This highlights the importance of better understanding the role of the informal sector in the design of policy.

The study is organized as follows. In Section 3.2, we present the baseline model and its implications for hazard rates from the informal to the formal sector. In Section 3.3, we present models with human capital accumulation and with employer learning, deriving their implications for hazard rates. Once the theoretical implications are described, in Section 3.4 we describe the data used in the empirical analysis. The details of the estimation follow in Section 3.5. Section 3.6 summarizes the empirical results, and Section 3.7 concludes

[4] In this stream of the program, the firm can participate in the selection and recruitment of workers participating in the program.

with some remarks on the results and suggestions for future research.

3.2 Baseline Model

The labor market is composed of two sectors, a formal sector and an informal sector. Formal sector firms comply with labor regulations represented by a firing cost incurred by firms when jobs are destroyed. The firing cost is assumed to be a wasteful tax as in Mortensen and Pissarides (2003) and Dolado, Jansen, and Jimeno (2005), so no transfer to the worker takes place. Informal sector firms do not comply with labor regulations.

We follow Albrecht, Navarro, and Vroman (2006, 2009) by assuming that workers differ in their productivity in the formal sector, but they are equally productive in the informal sector. Workers in the formal sector produce px units per period, where $p \in \{p_L, p_H\}$, with $p_H > p_L$, and x is a measure of *match quality*. Match quality is a random draw from a known distribution $G(x)$ with support on $[0, 1]$ that is made when the worker and firm meet; match quality stays constant until the job is destroyed. A fraction ϕ of the workers have the innate productivity p_L in the formal sector; we refer to these workers as L-skilled and the others as H-skilled. Innate productivity is perfectly observable. All workers in the informal sector produce p_I units per period. It is assumed that $p_I \geq z$, where z is the flow utility in unemployment.

Job destruction in both sectors follows from an idiosyncratic shock that arrives to occupied jobs at Poisson rate δ. If the job is destroyed in the formal sector, the firm incurs a firing cost D. Jobs are also destroyed due to worker's death. A worker dies with probability τ regardless of the worker's employment status. Every dead worker is replaced by a new unemployed worker who is L-skilled with probability ϕ. Job destructions due to death do not generate firing costs.

Unemployed workers search for jobs in both sectors, and all informal sector workers search for jobs in the formal sector.[5] The number of meetings between workers and firms in the informal sector is $m(u, v_I)$ and $m(u + e_I, v_F)$ in the formal sector, where u and e_I are the number of workers in unemployment and in informal sector jobs, respectively, v_j is the number of open vacancies in sector $j \in \{F, I\}$, and $m(\cdot, \cdot)$ is the meeting function. The meeting function is homogeneous of degree one, concave and increasing in both its arguments. As a result, a job seeker meets a firm in sector $j \in \{F, I\}$ with probability $m(\theta_j) = m(1, \theta_j)$, and a firm in sector j meets a job seeker with probability $m(\theta_j)/\theta_j$, where $\theta_I = v_I/u$ and $\theta_F = v_F/(u + e_I)$ are the measures of market tightness in the informal and the formal labor markets, respectively.

Given the assumptions on productivity in the informal sector, all meetings between an informal sector firm and an unemployed worker lead to job creation. Due to firing costs and to the assumptions on productivity in the formal

[5]To focus on flows from the informal to the formal sector, we abstract from on-the-job search in the opposite direction and from on-the-job search within each sector.

sector, a job in this sector is created if and only if the match quality is higher than a reservation match quality. The reservation match quality is endogenous and depends on both the skill level and the current employment status of the worker.[6]

The payoffs for workers are:

$$\tilde{r}U(p) = z + m(\theta_I)\big[W_I(p) - U(p)\big] + m(\theta_F)\int_{C(p)}^{1}\big[W_F(s,p) - U(p)\big]dG(s) \tag{3.1}$$

$$\tilde{r}W_F(x,p) = w_F(x,p) + \delta\big[U(p) - W_F(x,p)\big] \tag{3.2}$$

$$\tilde{r}W_I(p) = w_I(p) + \delta\big[U(p) - W_I(p)\big] + m(\theta_F)\int_{Q(p)}^{1}\big[W_F(s,p) - W_I(p)\big]dG(s) \tag{3.3}$$

where $\tilde{r} \equiv r + \tau$ and r is the discount rate. $U(p)$, $W_F(x,p)$, and $W_I(p)$ denote the present discounted value of the expected income stream of an unemployed worker, a worker employed in the formal sector, and a worker employed in the informal sector, respectively. Employed workers earn wage $w_I(p)$ or $w_F(x,p)$ when they work in the informal or the formal sector, respectively. The reservation match quality for the unemployed is $C(p)$ and for informal sector workers is $Q(p)$.

For workers of skill level p, the value of unemployment, $\tilde{r}U(p)$, depends on three main factors. First, unemployed workers receive flow utility z, which can be thought as the utility derived from leisure. Second, if they find an informal sector job, they experience a gain of $[W_I(p) - U(p)]$, and this happens with probability $m(\theta_I)$. Third, if they find a formal sector job, they experience a gain

[6]We follow Dolado, Jansen, and Jimeno (2005) in this job creation mechanism with two worker skill levels, firing cost, and initial random draw determining cut-offs for job creation.

of $[W_F(s,p) - U(p)]$. For unemployed workers, the probability of finding a formal sector job depends on: (i) the probability of meeting a formal sector firm with an empty vacancy, $m(\theta_F)$, and (ii) the probability that the match is worth forming, i.e. that the match quality randomly drawn is higher than $C(p)$.

For formal sector workers of skill level p currently employed with match quality x, the value of formal sector employment, $\tilde{r}W_F(x,p)$, depends on two main factors. First, they receive wage $w_F(x,p)$. Second, with probability δ they lose their job and experience a loss of $[U(p) - W_F(x,p)]$.

Finally, for informal sector workers of skill level p, the value of informal sector employment, $\tilde{r}W_I(p)$, depends on three main factors. First, they receive wage $w_I(p)$. Second, with probability δ they lose their job and experience a loss of $[U(p) - W_I(p)]$. Third, if they find a formal sector job, they experience a gain of $[W_F(s,p) - W_I(p)]$. For informal sector workers, the probability of finding a formal sector job depends on: (i) the probability of meeting a formal sector firm with an empty vacancy, $m(\theta_F)$, and (ii) the probability that the match is worth forming, i.e. that the match quality randomly drawn is higher than $Q(p)$.

The payoffs for firms are:

$$\tilde{r}J_F(x,p) = px - w_F(x,p) + \delta\Big[V_F - D - J_F(x,p)\Big] + \tau V_F \tag{3.4}$$

$$\tilde{r}J_I(p) = p_I - w_I(p) + \big[\delta + \mu(p)\big]\Big[V_I - J_I(p)\Big] + \tau V_I \tag{3.5}$$

$$rV_F = -k_F + \frac{m(\theta_F)}{\theta_F}\Big(E_{X,P}\big[J_F(x,p)|\phi_U,\phi_I\big] - V_F\Big) \tag{3.6}$$

$$rV_I = -k_I + \frac{m(\theta_I)}{\theta_I}\Big(E_P\big[J_I(p)|\phi_U\big] - V_I\Big) \tag{3.7}$$

where $\mu(p) \equiv m(\theta_F)\big[1 - G\big(Q(p)\big)\big]$, $J_F(x,p)$, and $J_I(p)$ denote the present discounted value of the expected profit from an occupied job in the formal and the informal sector, respectively, and V_j denotes the present discounted value of expected profit from a vacant job in sector $j \in \{F, I\}$. Note that (3.4) incorporates firing costs, (3.5) incorporates the possibility that the worker moves to the formal sector, and that the value of an open vacancy depends on the recruitment costs, k_j, and on the fraction of low-skilled job seekers, given by ϕ_U in unemployment and ϕ_I in the informal sector.

For a firm in the formal sector matched with a worker of skill level p and current match quality x, the value of the filled vacancy, $\tilde{r}J_F(x,p)$, depends on three main factors. First, the firm has a profit of $[px - w_F(x,p)]$. Second, if the job is destroyed, the firm experiences a loss of $[V_F - D - J_F(x,p)]$. Third, if the worker dies, the firm is left with an empty vacancy, and this happens with probability τ.

For a firm in the informal sector matched with a worker of skill level p, the value of the filled vacancy, $\tilde{r}J_I(p)$, depends on three main factors. First, the firm has a profit of $[p_I - w_I(p)]$. Second, with probability δ the jobs is destroyed, and with probability $\mu(p)$ the worker quits in order to take a formal sector job. In both cases, the firm suffers a loss of $[V_I - J_I(p)]$. Third, if the worker dies, the firm is left with an empty vacancy, and this happens with probability τ.

In both sectors, the value of an open vacancy depends on two main factors. First, it depends on the recruitment costs, k_j, for $j \in \{F, I\}$. Second, it depends

on the gain from filling the vacancy. This gain depends on the distribution of L-skilled and H-skilled workers that may contact the firm. For informal sector firms, since only unemployed workers contact them, the gain is given by $[E_P\big[J_I(p)|\phi_U\big] - V_I]$, where $E_P\big[J_I(p)|\phi_U\big] = \phi_U J_I(p_L) + (1-\phi_U)J_I(p_H)$. For formal sector firms, since both unemployed and informal sector workers contact them, the gain is given by $[E_{X,P}\big[J_F(x,p)|\phi_U,\phi_I\big] - V_F]$, where:

$$E_{X,P}[J_F(x,p)|\phi_U,\phi_I] = \phi_U \int_{C(p_L)}^{1} J_F(x,p_L)dG(x) + (1-\phi_U)\int_{C(p_H)}^{1} J_F(x,p_H)dG(x) + \phi_I \int_{Q(p_L)}^{1} J_F(x,p_L)dG(x) + (1-\phi_I)\int_{Q(p_H)}^{1} J_F(x,p_H)dG(x),$$

where ϕ_U and ϕ_I are the steady state proportion of L-skilled workers in unemployment and in the informal sector, respectively.

Wages in both sectors are determined according to a surplus sharing rule that entitles workers to a fraction β of the match surplus. The match surplus in the informal sector is $S_I(p) = W_I(p) - U(p) + J_I(p) - V_I$, and in the formal sector is given by $S_F(x,p) = W_F(x,p) - U(p) + J_F(x,p) - V_F$. The resulting wages are presented in Appendix B.1.

The decision to create a job in the formal sector depends on the match quality drawn when the worker and the firm meet. If the firm meets with an unemployed worker, both the firm and the worker require $x \geq C(p)$ to match, where $C(p)$ is such that $S_F(C(p),p) = 0$ for $p \in \{p_L, p_H\}$. If the firm meets with a worker in the informal sector, they require $x \geq Q(p)$, where $Q(p)$ is such that $S_F(Q(p),p) = S_I(p)$ for $p \in \{p_L, p_H\}$. Using the payoffs and wages, these cut-offs

are given by:

$$C(p) = \frac{\tilde{r}U(p)}{p} + \frac{\delta D}{p} \tag{3.8}$$

$$Q(p) = C(p) + \frac{(\tilde{r}+\delta)S_I(p)}{p} \tag{3.9}$$

where $p \in \{p_H, p_L\}$. Note that from (3.8) and (3.9) we cannot determine if $C(p_H) < C(p_L)$ and $Q(p_H) < Q(p_L)$ without some assumptions on productivity levels in the formal and informal sectors. Lemma 1 provides a sufficient condition that enables us to determine the relative size of the cut-offs.

Lemma 1. *Let $g(x)$ be the probability density function of the random variable x with support on $[0,1]$ representing match quality. Let $\eta = \frac{1}{1-\beta}\left(\frac{p_L}{p_I - z}\right)$. If*

$$\forall x \in [0,1] \quad \left(g(x) - \eta \int_x^1 g(u)du\right) < \eta(\tilde{r}+\delta),$$

then $C(p_H) < C(p_L)$ and $Q(p_H) < Q(p_L)$.

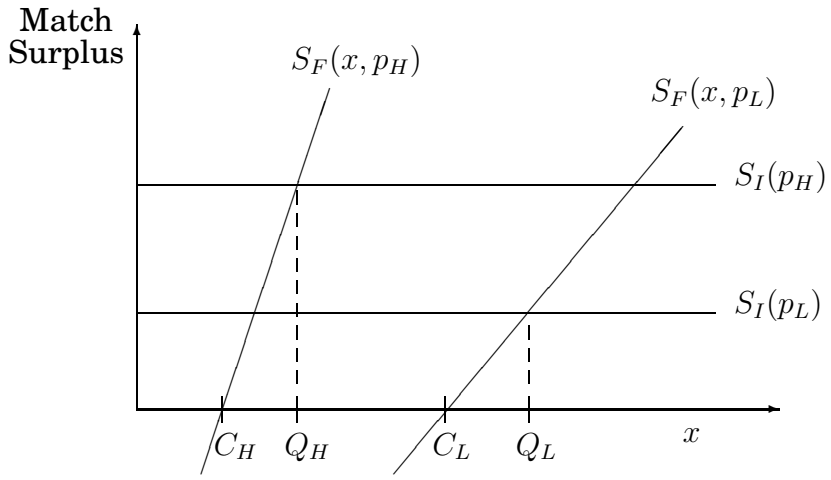
Appendix B.2.1 presents the proof of Lemma 1. The condition in Lemma 1 is easily satisfied.[7] This condition requires the distribution of match quality to be smooth and without spikes, so that the random draw taken when the worker and firm meet is relevant in the decision to create a job or keep looking for a better match.

After substituting wages and cut-offs in the match surplus in the formal sector, we find that $S_F(x,p) = \frac{p}{\tilde{r}+\delta}(x - C(p))$. Then, given the result in Lemma 1 and that $p_H > p_L$, it follows that $\forall x \in [0,1]$, $S_F(x,p_H) > S_F(x,p_L)$ and

[7] Notice that if $p_L > p_I$, then $\eta > 1$, since by assumption $p_I \geq z$. The larger η, the easier for the condition in Lemma 1 to be satisfied.

$\partial S_F(x, p_H)/\partial x > \partial S_F(x, p_L)/\partial x$. Figure 3.4 illustrates this result, and the fact that $C(p_H) < C(p_L)$ and $Q(p_H) < Q(p_L)$. Note that $S_I(p) > 0$ implies that $Q(p) > C(p)$ for $p \in \{p_L, p_H\}$, as a consequence informal sector workers are more selective than unemployed workers when it comes to matching with a formal sector firm.

Figure 3.4: Reservation Match Quality for Employed and Unemployed Workers



NOTE: $C_H = C(p_H)$, $C_L = C(p_L)$, and $Q_H = Q(p_H)$, $Q_L = Q(p_L)$

The baseline model produces implications for the hazard rate from the informal to the formal sector. We distinguish between the hazard rate conditional on worker skill level, denoted $\lambda(t|p)$, and the unconditional (or average) hazard rate, denoted $\lambda(t)$; where t is the realization of a random variable $T \geq 0$ measuring duration of employment in the informal sector and $p \in \{p_L, p_H\}$. These results are summarized in Propositions 1 and 2.

Proposition 1. *Suppose that the condition in Lemma 1 holds. Then, in the*

baseline model, the hazard rate from the informal to the formal sector conditional on the worker skill level, $\lambda(t|p)$, is constant for each $p \in \{p_L, p_H\}$, and it is higher for H-skilled workers than for L-skilled workers.

Proof. In the baseline model, the hazard rate conditional on worker skill is given by $\lambda(t|p) = \mu(p) = m(\theta_F)\big[1 - G(Q(p))\big]$, so that $\partial\lambda(t|p)/\partial t = 0$. By Lemma 1, $Q(p_H) < Q(p_L)$, which implies that $\lambda(t|p_H) > \lambda(t|p_L)$. □

Proposition 2. *In the baseline model, the unconditional hazard rate, $\lambda(t)$, is decreasing in duration.*

The proof of Proposition 2 follows the arguments of Lancaster (1990) and is presented in Appendix B.2.2. In this model, the fraction of L-skilled workers in the *risk set* (i.e. those that have not left the informal sector yet) increases with duration, pushing down the average hazard rate. This fraction increases with duration because H-skilled workers move from the informal to the formal sector at a faster rate than L-skilled workers. Lancaster (1990) calls this a "selection effect."

3.3 Extensions to the Baseline Model

The baseline model provides an analytical framework that helps us understand the key factors underlying the transitions from the informal to the formal sector. However, this model predicts that the transition rates from the informal to the formal sector remain constant as workers age. Yet, as shown in Figures

3.2 and 3.3, this is not the case in the data. Instead, we observe that transition rates increase as workers age (during early stages of the workers' careers).

We consider two extensions to the baseline model intended to explain this feature in the data. First, we assume that workers can accumulate human capital while working, which increases the chance of finding a formal sector job. Second, we assume that employers gradually learn about workers' skills. As a result, workers who are found to be H-skilled increase their chances of finding a formal sector job. We implement each extension separately because, as shown below, each mechanism generates opposing implications that would be hard to disentangle in a model with both mechanisms.

We focus on the implications for the hazard rate from the informal to the formal sector. On the one hand, when we assume that a worker can become more productive while in the informal sector, the longer such a worker stays in this sector, the more likely he is to make a transition into the formal sector. On the other hand, when we assume that a worker's productivity is gradually learned, those discovered as highly productive move to the formal sector faster, leaving behind those with low productivity levels and hence greater difficulties to access formal sector jobs. Thus, the longer a worker stays in the informal sector, the lower the likelihood that he makes a transition to the formal sector.

3.3.1 Human Capital Accumulation

First, we extend the baseline model by adding the possibility that workers accumulate skills through learning-by-doing. We follow Rebière (2008) and assume that a L-skilled worker can accumulate skills and become H-skilled with probability κ.[8] The accumulation of skills can only take place on the job, so the unemployed L-skilled workers cannot become H-skilled. Human capital does not depreciate, but since workers die and are replaced, the model does not converge to a degenerate distribution of skills.

The payoffs for unemployed workers and for vacancies have the same formulation as in the baseline model. The payoffs for employed workers and for filled vacancies now incorporate the possibility of accumulating skills. These are given by:

$$\tilde{r}W_F(x,p) = w_F(x,p) + \delta\big[U(p) - W_F(x,p)\big] + \kappa\big[W_F(x,p_H) - W_F(x,p)\big] \tag{3.10}$$

$$\tilde{r}W_I(p) = w_I(p) + \delta\big[U(p) - W_I(p)\big] + \kappa\big[W_I(p_H) - W_I(p)\big] + m(\theta_F)\int_{Q(p)}^{1}\big[W_F(s,p) - W_I(p)\big]dG(s) \tag{3.11}$$

$$\tilde{r}J_F(x,p) = px - w_F(x,p) + \delta\big[V_F - D - J_F(x,p)\big] + \kappa\big[J_F(x,p_H) - J_F(x,p)\big] + \tau V_F \tag{3.12}$$

$$\tilde{r}J_I(p) = p_I - w_I(p) + \big[\delta + \mu(p)\big]\Big[V_I - J_I(p)\Big] + \kappa\Big[J_I(p_H) - J_I(p)\Big] + \tau V_I. \tag{3.13}$$

[8] In Rebière (2008), workers start as beginners and become experienced while working in the beginners' sub-market; once they are experienced they search for jobs in the experienced sub-market. The labor market is segmented, so only beginners search for jobs in the beginners' sub-market, and only experienced search for jobs in the experienced sub-market.

The terms that account for the accumulation of skills disappear when $p = p_H$, so the value functions for H-skilled workers have the same formulation as in the baseline model.

For L-skilled workers in this model, the value of employment increases by the possibility of accumulating human capital, which happens with probability κ either in the formal or in the informal sector. The worker's gain from human capital accumulation is given by $[W_F(x, p_H) - W_F(x, p_L)]$ if the worker is employed in the formal sector, and by $[W_I(p_H) - W_I(p_L)]$ if the worker is employed in the informal sector.

Firms with filled vacancies also benefit from the worker's human capital accumulation. A formal sector firm matched with a L-skilled worker and current match quality x experiences a gain of $[J_F(x, p_H) - J_F(x, p_L)]$ with probability κ, and an informal sector firm matched with a L-skilled worker experiences a gain of $[J_I(p_H) - J_I(p_L)]$ with probability κ.

Wages are determined by the surplus sharing rule. The resulting wages for this model are presented in Appendix B.1. The reservation match qualities for unemployed and employed workers are determined in terms of the match surplus in the formal sector. That is, $S_F(C(p), p) = 0$ and $S_F(Q(p), p) = S_I(p)$. In this model the cut-offs are given by:

$$C(p) = \frac{\tilde{r}U(p)}{p} + \frac{\delta D}{p} - \kappa\left(\frac{U(p_H) - U(p)}{p}\right) - \kappa\left(\frac{S_F(C(p), p_H)}{p}\right) \tag{3.14}$$

$$Q(p) = C(p) + \frac{(\tilde{r} + \delta)S_I(p)}{p} - \kappa\left(\frac{S_F(Q(p), p_H) - S_I(p) - S_F(C(p), p_H)}{p}\right) \tag{3.15}$$

where the terms that account for the accumulation of skills disappear when $p = p_H$. Note that the direct effect of human capital accumulation is to reduce the cut-offs for L-skilled workers; this effect is picked up by the negative terms in both (3.14) and (3.15). An indirect effect of human capital accumulation increases the cut-offs for L-skilled, because both the value of unemployment and the match surplus in the informal sector increase.

Obtaining results similar to those in Lemma 1 is much more complicated with the inclusion of human capital accumulation. Consider environments which satisfy the following conditions:

Condition 1. $\forall x \in [0,1]$, $S_F(x, p_H) > S_F(x, p_L)$.

Condition 2. $\forall x \in [0,1]$, $S_F(x, p_H) - S_F(x, p_L) > S_I(p_H) - S_I(p_L)$.

These two conditions impose complementarities between the production technology in the formal sector and worker skills. Condition 1 implies that formal sector firms have a strict preference for H-skilled workers. If satisfied, then $C(p_H) < C(p_L)$. Condition 2 implies that the marginal value of skills is higher in the formal sector than in the informal sector. If satisfied, then $Q(p_H) < Q(p_L)$. These two implications can be easily verified in Figure 3.4.

If Conditions 1 and 2 are satisfied, the human capital model preserves the same ranking in cut-offs as in the baseline model. With this, we can derive similar implications for the conditional and unconditional hazard rates. These results are summarized in Propositions 3 and 4.

Proposition 3. *Suppose that Conditions 1 and 2 are satisfied. Then, in the model with human capital accumulation, the hazard rate from the informal to the formal sector conditional on worker's initial skill level, $\lambda(t|p)$, is constant for H-skilled workers and increasing for L-skilled workers.*

Proof. The conditional hazard rate for H-skilled workers is $\lambda(t|p_H) = \mu(p_H)$, which is constant with respect to duration, t. Next, for L-skilled workers, the conditional hazard rate is given by $\lambda(t|p_L) = (1-\kappa)^t\mu(p_L) + [1-(1-\kappa)^t]\mu(p_H)$. Then: $\partial\lambda(t|p)/\partial t = (1-\kappa)^t \ln(1-\kappa)\big[\mu(p_L)-\mu(p_H)\big] > 0$, which is positive because $\mu(p_L) < \mu(p_H)$ and $\kappa \in (0,1)$. □

When workers accumulate skills while working in the informal sector, the increase in productivity derived from the accumulation of skills facilitates access to job opportunities in the formal sector. Consequently, the likelihood of moving from the informal to the formal sector for L-skilled workers increases with tenure in the informal sector, resulting in an increasing hazard for L-skilled workers.

Proposition 4. *Suppose that Conditions 1 and 2 are satisfied. Let ϕ_I be the probability that $p = p_L$ in the informal sector. Then, in the model with human capital accumulation, the unconditional hazard rate, $\lambda(t)$, is:*

(i) increasing if: $-\ln(1-\kappa) > (1-\phi_I)\big[\mu(p_H) - \mu(p_L)\big]$

(ii) U-shaped otherwise.

The proof of Proposition 4 follows the arguments of Lancaster (1990) and

is presented in Appendix B.2.2. This Proposition states that when κ is large, the higher transition rate to the formal sector of H-skilled workers does not increase the fraction of L-skilled in the risk set, because L-skilled workers accumulate skills at a faster rate. As such, the hazard rate is increasing in duration. In contrast, if κ is not very large, it takes some time for the L-skilled to accumulate skills, and the higher transition rate of the H-skilled results in a higher fraction of L-skilled in the risk set. In this case, the hazard rate is initially decreasing. However, eventually L-skilled workers accumulate skills, so the fraction of L-skilled in the risk set decreases, resulting in an increasing hazard for higher durations.

3.3.2 Employer Learning (Screening)

In this extension of the baseline model, we abstract from human capital accumulation. Instead, we assume that when workers enter the labor market, their skill level (or type) is not known, but it is eventually revealed while they are working. We refer to these workers as "newcomers." We assume that neither the worker nor the employer knows the newcomer's type, and that once the type is revealed, everybody can observe the worker's skill level, as in Farber and Gibbons (1996). The revelation process is a stochastic process such that the worker's skill is revealed with probability σ.

All newcomers start unemployed, and it is common knowledge that a fraction ϕ of them are L-skilled. Newcomers also follow a reservation match quality strategy when facing formal sector job opportunities, taking informal sector opportunities as they arrive. When the worker's type is revealed in a formal sector job, the job could be destroyed if the current match quality is below the reservation match quality for that worker's type.

Let C be the reservation match quality for unemployed newcomers, and Q be the reservation match quality for newcomers holding an informal sector job. In the current study we focus on cases that satisfy the following condition:

Condition 3. $C(p_H) < C < C(p_L)$ ***and*** $Q(p_H) < Q < Q(p_L)$.

If Condition 3 holds, then all formal sector workers found to be H-skilled keep their job. On the contrary, a formal sector worker found to be L-skilled with match quality $x < C(p_L)$ loses his job, in which case the firm incurs firing costs. If the worker is found to be L-skilled but match quality is $x > C(p_L)$, then the worker keeps his job.

The payoffs and the reservation match quality for L-skilled and H-skilled workers have the same formulation as that in the baseline model. Let $\bar{p} \equiv \phi p_L + (1-\phi)p_H$ reflect the expected formal sector productivity for newcomers. Given Condition 3 holds, the payoffs for newcomers are given by:

$$\tilde{r}U = z + m(\theta_I)[W_I - U] + m(\theta_F)\int_C^1 [W_F(s) - U]dG(s) \tag{3.16}$$

$$\tilde{r}W_F(x) = w_F(x) + \delta[U - W_F(x)] + \sigma(1-\phi)W_F(x, p_H) + \sigma\phi\Big[\Gamma_L(x)U(p_L) + (1-\Gamma_L(x))W_F(x, p_L)\Big] - \sigma W_F(x) \quad (3.17)$$

$$\tilde{r}W_I = w_I + \delta[U - W_I] + m(\theta_F)\int_Q^1 [W_F(s) - W_I]dG(s) + \sigma\phi W_I(p_L) + \sigma(1-\phi)W_I(p_H) - \sigma W_I \quad (3.18)$$

$$\tilde{r}J_F(x) = \bar{p}x - w_F(x) + \delta[V_F - D - J_F(x)] + \sigma(1-\phi)J_F(x, p_H) + \sigma\phi\Big(\Gamma_L(x)[V_F - D] + (1-\Gamma_L(x))J_F(x, p_L)\Big) - \sigma J_F(x) + \tau V_F \quad (3.19)$$

$$\tilde{r}J_I = p_I - w_I + [\delta + \bar{\mu}][V_I - J_I] + \sigma\phi J_I(p_L) + \sigma(1-\phi)J_I(p_H) - \sigma J_I + \tau V_I \quad (3.20)$$

where $\bar{\mu} \equiv m(\theta_F)[1 - G(Q)]$, and $\Gamma_L(x) = \mathbf{1}\{x < C(p_L)\}$.

For unemployed newcomers, the value of unemployment, $\tilde{r}U$, depends on three main factors. First, unemployed newcomers receive flow utility z. Second, if they find an informal sector job, they experience a gain of $[W_I - U]$, and this happens with probability $m(\theta_I)$. Third, if they find a formal sector job, they experience a gain of $[W_F(s) - U]$. For unemployed newcomers, the probability of finding a formal sector job depends on: (i) the probability of meeting a formal sector firm with an empty vacancy, $m(\theta_F)$, and (ii) the probability that the match is worth forming, i.e. that the match quality randomly drawn is higher than C.

For newcomers employed in the formal sector with current match quality x, the value of formal sector employment, $W_F(x)$, depends on three main factors. First, they receive wage $w_F(x)$. Second, they experience a loss of $[U - W_F(x)]$ if

the job is destroyed, which happens with probability δ. Third, with probability σ, their skill level is revealed, in which case they might keep or lose their job. On the one hand, if a worker is found to be H-skilled, then the worker keeps his job and experiences a gain of $[W_F(x, p_H) - W_F(x)]$. On the other hand, if the worker is found to be L-skilled, then two things may happen: (i) if the current match quality is higher than $C(p_L)$, then the worker keeps his job, but experiences a loss of $[W_F(x, p_L) - W_F(x)]$, and (ii) if the current match quality is lower than $C(p_L)$, then the worker loses his job and experiences a loss of $[U(p_L) - W_F(x)]$.

By comparison, for formal sector firms matched with a newcomer with current match quality x, the value of the filled vacancy, $\tilde{r}J_F(x)$, depends on three main factors. First, the firm has profit $[\bar{p}x - w_F(x)]$. Second, with probability δ the job is destroyed and the firm suffers a loss of $[V_F - D - J_F(x)]$. Third, with probability σ the worker's skill level is revealed, and then the firm may keep the worker or let him go. If the worker is found to be H-skilled, the firm keeps the worker, and experiences a gain of $[J_F(x, p_H) - J_F(x)]$. If the worker is found to be L-skilled, two things can happen: (i) if the current match quality is higher than $C(p_L)$, then the firm keeps the worker, but suffers a loss of $[J_F(x, p_L) - J_F(x)]$, and (ii) if the current match quality is lower than $C(p_L)$, then the firm has to let the worker go, suffering a loss of $[V_F - D - J_F(x)]$.

For newcomers employed in the informal sector, the value of informal sector employment, $\tilde{r}W_I$, depends on four main factors. First, newcomers receive

wage w_I. Second, if the job is destroyed, they suffer a loss of $[U - W_I]$. Third, if they find a formal sector job, they experience a gain of $[W_F(s) - W_I]$. For newcomers employed in the informal sector, the probability of finding a formal sector job depends on: (i) the probability of meeting a formal sector firm with an empty vacancy, $m(\theta_F)$, and (ii) the probability that the match is worth forming, i.e. that the match quality randomly drawn is higher than Q. Fourth, if their skill level is revealed, then they experience a gain of $[W_I(p_H) - W_I]$ if they are found to be H-skilled, and a loss of $[W_I(p_L) - W_I]$ if they are found to be L-skilled.

By comparison, for informal sector firms matched with a newcomer, the value of the filled vacancy, $\tilde{r}J_I$, depends on four main factors. First, the firm has profit $[p_I - w_I]$. Second, with probability δ the job is destroyed. Third, with probability $\bar{\mu}$ the worker quits to take a formal sector job. In any of these two situations, the firm suffers a loss of $[V_I - J_I]$. Fourth, if the worker skill level is revealed, then the firm experiences a gain of $[J_I(p_H) - J_I]$ if the worker is found to be H-skilled, and a loss of $[J_I(p_H) - J_I]$ if the worker is found to be L-skilled.

Wages for this model are presented in Appendix B.1. Given Condition 3, reservation match qualities for newcomers are:

$$C = \frac{\tilde{r}U}{\bar{p}} + \frac{\delta D}{\bar{p}} + \frac{\sigma\phi D}{\bar{p}} - \frac{\sigma[\phi U(p_L) + (1-\phi)U(p_H) - U]}{\bar{p}} - \frac{\sigma(1-\phi)S_F(C, p_H)}{\bar{p}} \quad (3.21)$$

$$Q = \frac{\tilde{r}U}{\bar{p}} + \frac{\delta D}{\bar{p}} + \frac{\Gamma_L(Q)\sigma\phi D}{\bar{p}} - \frac{\sigma[\phi U(p_L) + (1-\phi)U(p_H) - U]}{\bar{p}} - \frac{[1-\Gamma_L(Q)]\sigma\phi S_F(C, p_L)}{\bar{p}} - \frac{\sigma(1-\phi)S_F(Q, p_H)}{\bar{p}} + \frac{(\tilde{r} + \delta + \sigma)S_I}{\bar{p}}. \quad (3.22)$$

Note that if $\Gamma_L(Q) = 1$, then $Q \approx C + \dfrac{(\tilde{r} + \delta + \sigma)S_I}{\bar{p}}$. Again, these hiring standards give us some implications in terms of the hazard rates from the informal to the formal sector, which are summarized in Propositions 5 and 6.

Proposition 5. *Suppose that the condition in Lemma 1 and Condition 3 hold. Then, in the model with employer learning, the hazard rate from the informal to the formal sector conditional on the worker skill level, $\lambda(t|p)$, is increasing for H-skilled workers and decreasing for L-skilled workers.*

Proof. The conditional hazard rate is given by $\lambda(t|p) = (1-\sigma)^t\bar{\mu}+[1-(1-\sigma)^t]\mu(p)$, for each $p \in \{p_H, p_L\}$. Let $\partial\lambda(t|p)/\partial t = \lambda'(t|p)$, then $\lambda'(t|p) = (1-\sigma)^t \ln(1-\sigma)\big[\bar{\mu} - \mu(p)\big]$, which is positive for $p = p_H$ because $\bar{\mu} < \mu(p_H)$ and $\sigma \in (0,1)$, and negative for $p = p_L$ because $\bar{\mu} > \mu(p_L)$ and $\sigma \in (0,1)$. □

In this model, employers can distinguish three different groups of workers. However, everyone knows that newcomers are either L-skilled or H-skilled. H-skilled workers face an increasing hazard in their informal sector career because once they are revealed as H-skilled, the likelihood of finding a formal sector job increases. On the contrary, L-skilled workers face a decreasing hazard.

Proposition 6. *Suppose that the condition in Lemma 1 and Condition 3 hold. Let ϕ be the probability that $p = p_L$ in the labor market. Then, in the model with employer learning, the unconditional hazard rate, $\lambda(t)$, is:*

(i) decreasing if $\bar{\mu} > \phi\mu(p_L) + (1-\phi)\mu(p_H)$

(ii) hump-shaped otherwise.

The proof of Proposition 6 follows the arguments of Lancaster (1990) and is presented in Appendix B.2.2. Proposition 6 states that the shape of the unconditional hazard function initially depends on whether the hazard rate of newcomers is higher or lower than the average hazard rate of workers with revealed types. Cases (i) and (ii) compare these two hazard rates. Eventually, as more worker types are revealed, the hazard function decreases with duration due to selection, as in the baseline model.

Whether case (i) or (ii) arises depends on: $a)$ the mixture of H-skilled and L-skilled workers in the population, summarized by ϕ; $b)$ the location of Q with respect to $Q(p_L)$ and $Q(p_H)$; and $c)$ the properties of the distribution of match quality, $G(x)$. Note that Q is not determined by $Q\big(\phi p_L + (1-\phi)p_H\big)$, and so we cannot raise conclusions in terms of the properties of $Q(\cdot)$, defined in equation (3.9). Even so, case (i) is more likely to occur if $\overline{G}(x) \equiv [1 - G(x)]$ is concave (or $G(x)$ convex), so that the convex combination $\phi\overline{G}\big(Q(p_L)\big) + (1-\phi)\overline{G}\big(Q(p_H)\big)$ is lower than $\overline{G}(Q)$. In contrast, case (ii) is more likely to arise if $\overline{G}(x)$ is convex (or $G(x)$ concave), so that the convex combination of $\overline{G}\big(Q(p_L)\big)$ and $\overline{G}\big(Q(p_H)\big)$ is higher than $\overline{G}(Q)$.[9]

[9]Simulation exercises assuming that $G(\cdot)$ is uniform indicate that whether case (i) or (ii) arises is mainly determined by the fraction of L-skilled workers in the population, ϕ. These exercises show that ϕ is the main determinant of the location of Q with respect to $Q(p_L)$ and $Q(p_H)$. The larger ϕ is, the closer Q is to $Q(p_L)$, and the more likely that case (ii) arises. Intuitively, when ϕ is large, formal sector employers treat "newcomers" as if they were L-skilled. As a result, both "newcomers" and L-skilled workers in the informal sector move to the formal sector at similar rates, whereas H-skilled in the informal sector move at faster rates. Hence, for short spells (low t) the hazard increases when the first "newcomers" see their

3.3.3 Understanding the Role of the Informal Sector in the Early Careers of Less-educated Workers

We are now in a position to assess the role of the informal sector in the early stages of the careers of less-educated workers. The implications derived earlier suggest estimating the hazard function from the informal to the formal sector to determine whether human capital accumulation or screening/learning are important in the informal sector. We estimate these hazard functions using data from an employment survey from Mexico. In the next section, we describe the data, the sample, and some details of the variables used in estimation.

3.4 Data: The ENOE

We use a household survey from Mexico called the Occupation and Employment Survey, ENOE (its acronym in Spanish). The ENOE is a rotating panel where households are visited five times during 12 months, one visit every three months. Every three months, 20% of the sample is replaced. Although information from each family member is recorded, this information is provided by only one member; the respondent is not necessarily the same individual on each visit.

The ENOE records the demographics of each family member (e.g. education,

skill level being revealed because H-skilled have a faster exit rate than both "newcomers" and L-skilled.

age, marital status), and information on the main and secondary jobs of family members older than 12 years of age. Job information includes working hours, earnings, fringe benefits, job position, firm size, industry, occupation and job tenure. The job tenure information is only recorded in the *long form* of the ENOE, which is answered at least once during the five visits to the household. For further details about the ENOE see INEGI (2005, 2007).

3.4.1 Sample

To focus on less-educated workers, we restrict the sample to individuals not currently attending school and with less than 12 years of education. To focus on young workers, our sample only includes workers between the ages of 16 and 25. Age 16 is the minimum age at which a worker can be hired according to Mexican Labor Law (see Congress, 1970), and age 25 is the age at which transitions from the informal to the formal sector plateau (see Figures 3.2 and 3.3). Our sample only includes male workers because women may have different reasons for joining the informal sector, e.g. job flexibility to balance work and child rearing (Arias and Maloney, 2007).

We divide our sample of less-educated workers into two groups based on completion of the mandatory level of education in Mexico, which is 9 years. In one group, we include less-educated workers who failed to complete the mandatory level of education, and in the other those who completed the mandatory level of education but who failed to complete high school (i.e. 12 years). Since

the mandatory level of education in Mexico could be compared to junior high school in the U.S., we refer to the first group as junior high school dropouts, and the second as junior high school graduates.[10]

Table 3.1 presents the sample summary statistics. For the purpose of this table, the group of junior high school dropouts is further divided in two groups. Junior high school graduates represent 63% of the sample. Workers in all three groups are mainly concentrated in small firms, but the junior high school graduates have the highest percentage in large firms. Also, note that the two groups of junior high school dropouts are mainly concentrated in the construction industry, while graduates are mainly concentrated in the services industry. Finally, note that graduates are more likely to have a parent working in a formal sector job. Firm size, industry, and family head employment status could be important determinants of the probability of moving from the informal to the formal sector.

3.4.2 Identification of Informal Salaried Workers

When a worker is hired in Mexico, it is the employer's responsibility to register the worker in the IMSS or the ISSSTE.[11] These institutions provide a bundle of benefits to their affiliates. For example, the bundle offered by IMSS includes:

[10] In Mexico, compulsory education comprises primary school (grades 1 to 6) and junior high school (grades 7 to 9). In terms of our labeling, note that some of the individuals in the junior high school dropout group may not have even started junior high school.

[11] IMSS is the acronym in Spanish for the Mexican Institute of Social Security and ISSSTE is the acronym in Spanish for the Institute of Security and Social Services for the State's Workers.

health insurance, day-care services for children, life insurance, disability pensions, work-risk pensions, sports and cultural facilities, retirement pensions, and housing loans (Levy, 2007). Both the worker and the employer must pay fees to fund these institutions, but the portion paid by the employer is much higher than that paid by the worker. If the firm is caught not complying with these regulations, it incurs a penalty.

Once a worker is registered in the IMSS or the ISSSTE the work relationship must abide by the labor regulation in Mexico. This means that the employer will incur firing costs if the work relationship is terminated.

The questionnaire of the ENOE does not ask the individual whether he is a formal or an informal worker. Instead, the survey asks the individual if his job gives him access to medical services provided by: the IMSS, the ISSSTE, the military hospital, the PEMEX hospital, or any other hospital (i.e. private hospital).[12] We consider a worker to belong to the formal sector if he is an employee and his job gives him access to any kind of medical services: from IMSS, ISSSTE, military, PEMEX, or private; and to belong to the informal sector if he is an employee and his job does not give him access to any of these services. Note that the self-employed are not included in our definition of the informal sector.

[12]PEMEX is the state-owned petroleum company in Mexico. Both, military and PEMEX workers, have access to medical services independent of IMSS or ISSSTE. Workers that have access to private medical services are usually hired formally, and even though they do not use the medical service of IMSS, they are registered at the IMSS, and could use it if desired.

3.4.3 Measuring Duration in the Informal Sector

Duration of employment in the informal sector is obtained using two different sampling schemes: flow sampling and stock sampling. In the flow sample, we include individuals who enter the informal sector during a fixed period of time, namely the 12 months in which the ENOE follows households.[13] In the stock sample, we include individuals who are already in the informal sector at a given point in time, namely the month of the visit in which the household answered the *long form* of the ENOE. The date of the visit in which the *long form* is answered is used as the stock sampling date because the *long form* records the starting date of the current job.[14] The starting date is either recorded as: (i) the exact month, if the job started in the current or the previous calendar year, or (ii) the year, if the job started before the previous calendar year.

Duration of employment in the informal sector is defined as the length of time that passes between the point in time in which the respondent enters the informal sector and the point in time in which the respondent moves from the informal to the formal sector. Duration is right-censored if the respondent is still employed in the informal sector at the time of the last interview. Duration of employment for individuals who leave the informal sector but do not enter the formal sector is also treated as right-censored.

[13] We include in this sample individuals who enter the informal sector after the first but before the fourth visits. Those who made a transition between the fourth and fifth visits are not included, because we are not able to follow them after the fifth visit.

[14] We include in this sample informal sector workers whose *long form* interview took place in the first, second, third, or fourth visits. If an individual answered the *long form* for more than one visit, we use the first one as the stock sampling date.

Given that the household is visited every three months, the point in time of the transition from the informal to the formal sector either is known to be: (i) the exact month, or (ii) contained in a 3-month interval. The second case can arise in two situations: (i) if the respondent made such transition without changing jobs, or (ii) if the respondent changed jobs, but the visit to the household following that transition did not use the *long form* questionnaire, and so the starting time of the new job was not recorded.

Consequently, combining the two different formats in which the starting time and the transition time are recorded, duration from the stock sample is either known to the exact number of months or contained within some interval. On the other hand, all duration measures from the flow sample are interval-censored. This is because the starting time is never exactly known, but only known within three months (the time between interviews). Thus, whether the point in time of the transition is known to the month or within a 3-month interval does not change the fact that the completed duration will be only known within an interval.

Table 3.2 describes the distribution of formats in which duration in the informal sector is recorded in the sample. The most frequent intervals are 6-month for the flow sample, and 3-month and 15-month for the stock sample. This is a result of the frequency in which the household is visited.[15] The

[15]In the flow sample, both the point in time in which the individual enters the informal sector and the point in time in which the individual moves to the formal sector can be only known within a 3-month interval, which results in a 6-month interval. This turns out to be the most frequent case in the sample. In the flow sample, the starting time of the job is either know to

numbers in the table also reveal the fact that the sample is subject to a high degree of right-censoring. Sixty percent of the spells in the sample are right-censored. Table 3.3 describes the source of censoring in the sample. Among those censored observations, 52% are due to the respondent being employed in the informal sector at the time of the last interview, 29% are due to the respondent moving out of the informal sector to another work status, such as self-employment, and 20% are due to the respondent becoming unemployed.

To summarize, let T_i be the duration of employment in the informal sector for respondent i. We either observe T_i up to the exact number of months, or an interval $(L_i, R_i]$ such that $T_i \in (L_i, R_i]$. Similarly, let C_i be the censoring time for respondent i. Then, for censored observations we only know that $T_i > C_i$, or that $(L_i, R_i] = (C_i, \infty)$. Notice that because different respondents have different starting dates, the intervals $(L_i, R_i]$ may overlap for different respondents. As a result, we cannot use the techniques of discrete time duration analysis (e.g. Prentice and Gloeckler, 1978; Meyer, 1990; Han and Hausman, 1990). We must instead work with interval-censored data (e.g. Finkelstein, 1986; Sun, 2006).

Finally, some of the spells in the sample have starting times on a date before the individual reaches age 16. Individuals who started their informal sector jobs before age 16 may delay their transition to the formal sector owing to legislative restrictions, and not for the reasons stipulated in the model. We adjust

the month or within a 12-month interval, and the point in time of the transition to the formal sector can be known within a 3-month interval, which results in a 3-month or in a 15-month interval. These are the two most frequent cases in the sample.

the duration measure of these individuals by subtracting from their duration the number of months worked before age 16, and create an indicator variable for them, which is included in the covariates. In this way, all job spells measure the time that the individuals were "at risk" of making a transition to the formal sector. About 2% of the spells in the sample are adjusted because of their pre-age 16 starting point.

Table 3.4 summarizes the duration data generated from the ENOE. For this table, we impute interval-censored duration measures with the midpoint in the interval. Note that the mean duration of employment in the informal sector is lower for junior high school graduates. In fact, the distribution of duration for junior high school graduates first-order-stochastically dominates that of the dropouts, suggesting that graduates move to the formal sector at a faster rate than the dropouts.

Before proceeding with the estimation, it is important to mention that the implications from the models derived in the previous sections are in terms "sector spells." However, in the estimation below we will be using measures of "job spells." Given that we cannot follow the individual since the first time they entered the labor market, we have to work with the spell of the last job held by the individual. To the extent that the individual held other informal jobs before the current informal job, we would be underestimating the length of the sector spells, or in other words, the sector spells would be left-censored. It is in our advantage, however, that we are working with a sample of young workers, and

so we should expect that the job spells should be similar to the sector spells.

3.5 Estimation

3.5.1 Likelihood Function

The likelihood function is defined in terms of the hazard function, which is conditioned on a set of time-invariant covariates, x. The inclusion of covariates is very important in the presence of right-censoring in order to make valid inference. The right-censoring mechanism must satisfy the assumption of *independent censoring* (Kalbfleisch and Prentice, 1980). In terms of duration of employment in the informal sector, independent censoring requires that, conditional on x, an individual's duration is not censored because such individual has an unusually high (or low) probability of moving to the formal sector.[16]

In the ENOE, because all households are visited exactly five times, censoring as a result of the individual working in the informal sector during the last visit satisfies independent censoring. But we must be cautious with the duration of employment of individuals whose transition to the formal sector is not observed because they moved to another state (e.g. self-employment). The duration of employment of these individuals is right-censored, but the assumption

[16]Kalbfleisch and Prentice (1980) define a censoring scheme as independent if "the probability of censoring at time t depends only on the covariate x, the observed pattern of failures and censoring up to time t in the trial, or on random processes that are independent of the failure times in the trial." In the case of duration of employment in the informal sector, failure is defined as a transition from the informal to the formal sector.

of independent censoring could be violated if they were systematically more (or less) likely to make a transition to the formal sector.

To that end, in our covariates we include variables that also explain why these individuals move to another state *before* moving to the formal sector. The covariates include industry, firm size, educational attainment, government's financial support to self-employment, marital status, condition of employment of the family head, and dummies for different starting years.

As mentioned before, we use time-invariant covariates, although some of these covariates are in nature time-varying and could explain why some informal sector workers are more or less likely to make a transition to the formal sector. In particular, the covariates for marital status and firm size may vary over time and are important determinants of the transition from the informal to the formal sector. It is possible that during the time that the worker was employed in the informal sector his marital status changed from single to married, and so the increase in the demand for medical services associated with marriage could affect how fast the individual moves to the formal sector. Similarly, firms that are expanding might have an increased demand for formal sector jobs, or viceversa for a firm that is contracting. Levenson and Maloney (1998) find that, in the case of Mexico, firms treat formality as a "normal" input, and so its demand increases with the firm's expansion.

For the stock sample, we can only observe changes in the covariates after

the first interview, but we do not observe any changes before the first interview. For workers in the stock sample, we use the value of the covariate at the interview in which the long form of the ENOE was used. Although we can observe changes in the covariates for workers in the flow sample, in order to be consistent, we also fix the covariates for workers in this sample. For workers in the flow sample, we use the value of the covariate at the interview in which the informality status of the worker changed from not being informal to being an informal sector worker.

Interval-censoring also imposes a requirement in order to make inference, which is very similar to the one for right-censoring. Kalbfleisch and Prentice define this requirement as *independent interval censoring*. Let $0 < C_{i1} < C_{i2} < \cdots < C_{im_i} < \infty$ be the visiting dates for individual i. Independent interval censoring requires that: "having observed that the individual is [in the informal sector] at time $C_{i,j-1}$, the timing of the next [visit] is distributed independently of the time of the [transition to the formal sector]" (Kalbfleisch and Prentice, 1980, page 79). Since the household visits are scheduled every three months, this assumption is also satisfied in the ENOE. The assumption would be violated if the next visit is determined to be sooner (or later) depending on the probability that the individual moves from the informal to the formal sector.

Now, recall that 60% of the duration measures are from the stock sample (see Tables 3.2 and 3.3). It is known that stock sampling introduces a sample selection problem because long durations are more likely to be sampled

than short durations (Wooldridge, 2002). This problem, known as *length-biased sampling* (Kiefer, 1988), is easily addressed by including the starting time of the job spell in the likelihood, or more precisely the length of time that passes between the start of the job and the stock sampling date, which is know as the *elapsed duration*. Thus, in order to account for this sampling bias, we include the elapsed duration in the likelihood function.

Let T be a nonnegative random variable denoting the duration of employment in the informal sector, and let t be a particular value of T. Let $\lambda(t|x)$ be the hazard function of T, $S(t|x)$ be the survivor function, which is defined in terms of the hazard function as $S(t|x) = \exp\{-\int_0^t \lambda(s|x)ds\}$, and let $f(t|x)$ be the density of T, which is defined as $f(t|x) = \lambda(t|x)S(t|x)$. Given that both censoring mechanisms are independent, the contribution of a right-censored observation to the likelihood is $\Pr(T_i > C_i|x) = S(C_i|x)$, and the contribution of an interval-censored observation is $S(L_i|x_i) - S(R_i|x_i)$. Let e_i be the elapsed duration of individual i, then the likelihood function is given by:

$$L(\theta|x_i) = \prod_{\{i|\Upsilon_i=1\}} \frac{f(t_i|x_i)^{d_i} S(t_i|x_i)^{(1-d_i)}}{S(e_i|x_i)} \prod_{\{i|\Upsilon_i=0\}} \frac{S(L_i|x_i) - S(R_i|x_i)}{S(e_i|x_i)} \tag{3.23}$$

where Υ_i is an indicator for interval-censoring ($\Upsilon_i = 1$ if uncensored, $\Upsilon_i = 0$ if interval-censored), and d_i is an indicator for right-censoring ($d_i = 1$ if uncensored, $d_i = 0$ if right-censored).

Finally, as explained in the previous section, job starting times in the stock sample are either known up to the month, or up to the year. The likelihood function (3.23) assumes that we know e_i or, equivalently, that we know the

starting time. However, for some respondents, all we know is that this starting time is included in a 12-month interval, if the job started before the previous calendar year.

In order to overcome the coarseness of starting times, we performed a Monte Carlo analysis to explore different alternatives to impute the starting time when this information is interval-censored. The Monte Carlo analysis is presented in Chapter 4. Three methods to impute the elapsed duration were explored, using the: (i) lower bound of the interval, (ii) upper bound of the interval, and (iii) midpoint of the interval. The simulation results indicate that, for the case of duration data obtained from surveys like the ENOE, using the midpoint in the interval outperforms the alternatives. The empirical analysis in this paper follows the results from Chapter 4.

3.5.2 Hazard Function

To estimate the hazard, instead of imposing the functional form implied by each model, we estimate a flexible hazard function. Widely used parametric models such as the Weibull or the Log-logistic impose restrictions on the shape of the hazard (see Wooldridge, 2002, chap. 20). For this reason, our main results rely on the estimation of a piecewise constant hazard, which allows more flexibility in the shape of the hazard function. We assume a proportional hazards model $\lambda(t|x_i) = \exp(x_i'\rho)\lambda_0(t)$, where:

$$\lambda_0(t) = \lambda_m, \quad a_{m-1} \leq t < a_m, \quad \lambda_m > 0, \quad m = 1, 2, \ldots, M \tag{3.24}$$

and $\{a_0, a_1, \ldots, a_M\}$ are known break points that define $M+1$ intervals $[a_0, a_1)$, $[a_1, a_2)$, $\ldots$, $[a_{M-1}, a_M)$, $[a_M, \infty)$ that may contain t. We set $a_0 = 0$, and choose the other break points using the distribution of T. The distribution of T is divided into six quantiles, so that $M = 6$, with break points determined by the quantiles.[17]

The survivor function is given by:

$$S(t|x_i) = \exp\left\{ -\exp(x_i'\rho) \left[\sum_{k=1}^{I(t)-1} \lambda_k \Big(a_k - a_{k-1} \Big) + \lambda_{I(t)} \Big(t - a_{I(t)-1} \Big) \right] \right\} \tag{3.25}$$

where $I(t)$ is such that $a_{I(t)-1} \leq t < a_{I(t)}$, i.e. t is contained in the $I(t)^{th}$ interval.

We estimate the hazard function for the whole sample and for two mutually exclusive education groups. The break points for each of these samples are:

	months					
Education Group	a_1	a_2	a_3	a_4	a_5	a_6
$[\,0, 12\,)$	3.0	4.5	6.0	12.0	24.0	104.0
$[\,0, 9\,)$	3.0	4.5	7.0	12.0	24.0	104.0
$[\,9, 12\,)$	3.0	3.5	6.0	11.0	24.0	96.0

3.6 Results

3.6.1 Piecewise Constant Hazard Function

We maximize the likelihood function in equation (3.23) using all of the elements discussed in the previous section. The estimation results for the whole sample

[17] To avoid ties in the quantiles, the break points are the quantiles of $\hat{T}_i = (L_i + R_i)/2$.

and for junior high school dropouts and graduates are summarized in Table 3.5. Figure 3.5 depicts the estimated baseline hazard with the 95% pointwise confidence intervals. The plot of the baseline hazard in Figure 3.5 depicts the hump-shaped pattern predicted by the model with employer learning. Note that this pattern holds for the whole sample, and for the junior high school dropouts and graduates.[18]

Even though both junior high school dropouts and graduates show signs of employer learning, those who completed the mandatory level of education have a higher hazard rate at all times. In terms of Proposition 7, this result indicates that the proportion of L-skilled workers is higher among dropouts than among graduates as one might expect.[19]

Estimated effects of the covariates in Table 3.5 are fairly similar for the whole sample and for junior high school graduates and dropouts. The estimation results for the whole sample show that graduation from primary school (grade 6) has little effect on the hazard rates, but graduation from secondary school (grade 9) has a significant effect. This is consistent with Arias and Maloney (2007) who claim that "graduation to formal salaried work is unlikely for

[18]A similar estimation exercise was performed using only interval-censored weekly duration measures instead of using monthly duration measures. The estimation results indicate the same hump-shaped pattern in the hazard function for the three education groups.

[19]Alternatively, there could be more than two worker skill levels, with some of them concentrated in one education group, e.g. the highest concentrated in group of graduates and the lowest concentrated in the group of dropouts. Note that we could extend the models to a continuum of worker types, as in Albrecht, Navarro, and Vroman (2006, 2009). This would yield similar results to those derived above.

youth who drop out of school before completing at least a full course of secondary education" (Arias and Maloney, 2007, page 62) .

Not surprisingly, one of the most important covariates is the size of the firm. The higher the firm size, the higher the hazard rate from the informal to the formal sector. There are two potential explanations for this result. On the one hand, many of the transitions could be happening within the same employer. Alternatively, it could be that larger firms have a larger network and as a result expose workers' skills to other employers more than small firms do.

Industry does not play a big role in explaining the hazard rate from the informal to the formal sector. Married workers have higher hazard rates than single workers, consistent with the incremental demand for health services when individuals form their own families. And when the family head works in the formal sector, the individual also has a higher hazard rate, which could also be the result of the individual having access to a larger network of formal sector employers.[20] Notice that the estimates for these covariates are larger for junior high school dropouts than for the graduates.

Finally, note that a hump-shaped hazard rules out the baseline model. The baseline model predicts constant hazard rates conditional on worker skill level, which in turn implies that the unconditional survivor function is a mixture of exponential distributions. Based on comments made by Chamberlain (1980),

[20] In Mexico, dependents of workers registered in the IMSS can only use the medical services of this institution up to age 18. The coverage can be extended if the dependent is attending school, which is not the case in our sample.

Heckman, Robb, and Walker (1990) argue that "all mixtures of exponentials models have nonincreasing hazards." The pointwise confidence intervals for our estimated hazard imply that the hazard is increasing for short spells, thereby ruling out the baseline model (with any arbitrary number of worker types).[21]

3.6.2 Parametric Hazard Functions

As a robustness check, we estimated two widely used parametric hazards, the Weibull and the Log-logistic hazard models. We are mainly interested in the estimation result from the Log-logistic model. The Weibull is characterized by the hazard function:

$$\lambda(t) = \varphi\alpha t^{\alpha-1} \tag{3.26}$$

and the Log-logistic by:

$$\lambda(t) = \frac{\varphi\alpha t^{\alpha-1}}{1+\varphi t^{\alpha}}, \tag{3.27}$$

where $\varphi = \exp(x'\rho)$ is the most common choice in empirical applications. The shape of the hazard function in each case is determined by the parameter α, as summarized in the following table:

[21]Using the estimated hazard function, and following the procedure suggested by Chamberlain, we conclude that the survivor function for the data in this study cannot be generated by a mixture of exponentials. For a description of the rejection criterion and the procedure see Chamberlain (1980) or Heckman, Robb, and Walker (1990).

	Weibull	Log-logistic
$\alpha < 1$	Decreasing	Decreasing from ∞ at $t = 0$, to 0 as $t \to \infty$
$\alpha = 1$	Constant	Decreasing from φ at $t = 0$, to 0 as $t \to \infty$
$\alpha > 1$	Increasing	Increasing from 0 at $t = 0$, to a single maximum T^*, and then approaches 0 as $t \to \infty$

When $\alpha > 1$ in the Log-logistic, the maximum occurs at $T^* = [(\alpha - 1)/\varphi]^{1/\alpha}$ (see Lancaster, 1990, chap. 3).

The estimated hazards for these two models are presented in Table 3.6. The estimated coefficients for the covariates in the Weibull hazard are very similar to the ones in the piecewise constant hazard, since both of these are proportional hazards models. For the Log-logistic model, they are not identical but have the same pattern across the groups of dropouts and graduates from junior high school. Given the restrictions of the Weibull hazard, the estimates suggest a monotonically decreasing hazard, but the Log-logistic suggests a hump-shaped hazard. More importantly, the predicted maximum in the Log-logistic hazard function is very similar to the maximum we have in the piecewise constant hazard in Figure 3.5.[22]

[22]Likewise, a similar estimation exercise was performed using only interval-censored weekly duration measures, instead of using monthly duration measures. The estimation results yield very similar estimates for α in both the Weibull and the Log-logistic hazard functions.

3.6.3 Unobserved Heterogeneity

The models developed in the previous sections are based on the premise that there are two worker skill levels, which are not observable to the econometrician. For each model, we develop predictions based on the average (or unconditional) hazard function, which effectively integrates over any unobserved heterogeneity. In fact, changes in unobserved types over time play an important role in driving patterns of the average hazard functions used to identify the different models. An advantage of our approach is that it is based on the estimation of the average hazard function, and so it does not require us to specify a particular distribution for the unobserved heterogeneity.

3.6.4 A Final Comment on Testing the Implications

Notice that the implications of the three theoretical models presented in Propositions 2, 4, and 6, are defined in terms of duration of employment in the informal sector from the time the worker entered the labor market. However, the duration measures used in the estimation of the hazard functions are with respect to the last job of the individual. To the extent that the individual experienced previous job spells in the informal sector, we are underestimating this measure. On the other hand, the fact that we are working with a sample of young individuals suggests that the underestimation is not very severe. A similar estimation was performed using a younger sub-sample, which included

only individuals ages 16 to 20. The estimation results from this exercise are very similar and the hump-shape observed in the estimated hazard function persists. These results yield further support to the suggestion that underestimation of the duration of employment in the informal sector does not severely affect our results and conclusions.

3.6.5 Screening in *Bécate* Training Program

In this section we use the estimated piecewise constant hazard to infer the parameters governing the employer learning process. Knowledge of these parameters gives us the means to evaluate *Bécate*'s screening program introduced in Section 3.1. In terms of the employer learning model, we want to know how fast employers learn about their workers' abilities. This information is obtained using the model-generated hazard and the estimated hazard. The unconditional hazard in the employer learning model is a function of five parameters, $\big(\bar{\mu}, \mu(p_L), \mu(p_H), \sigma, \phi\big)$; while the piecewise constant hazard is a function of seven parameters, $(\lambda_1, \ldots, \lambda_6, \rho)$. We use the estimated parameters $\big(\hat{\lambda}_1, \ldots, \hat{\lambda}_6, \hat{\rho}\big)$ to infer the value of the parameters of the employer learning model.

Let $\nu(t) \equiv \lambda_M\big(t; \bar{\mu}, \mu(p_L), \mu(p_H), \sigma, \phi\big) - \lambda_{PW}\big(t; \hat{\lambda}_1, \ldots, \hat{\lambda}_6, \hat{\rho}\big)$, for $t = 0, 1, \ldots, \overline{T}$, denote the residual between the model generated hazard, $\lambda_M(\cdot)$, and the estimated piecewise constant hazard, $\lambda_{PW}(\cdot)$. To get the parameters governing the employers' learning process, we look for the vector $\big(\bar{\mu}, \mu(p_L), \mu(p_H), \sigma, \phi\big)$ that

minimizes the sum of squared residuals. The details of the optimization algorithm are explained in Appendix B.3.

The estimated and the model-generated hazards are shown in Figure 3.6. The resulting parameters indicate that employers learn their workers' abilities at a rate of $\sigma = 0.1478$ per month, and that the proportion of L-skilled workers in the population is $\phi = 0.4833$. Then, at the end of a three-month *Bécate* program, employers know the skill level of about 51% of the recruited workers, where 48% of these workers are expected to be L-skilled. The firm will be happy to hire those workers identified as H-skilled, but must also fulfill its promise to take 70% of the workers recruited for the program. This implies that the firm must take a gamble in hiring 44% of the original number of workers whose skill level is still unknown. However, since 48% of these workers are expected to to be L-skilled, the firm will end up hiring 21% of the original number of workers that are L-skilled. If the firm does not have a good match quality with these L-skilled workers, it will incur firing costs.

Note that the numbers we are getting from this exercise on the *Bécate* program are at the aggregate level. It must be the case that in some industries, the learning rate is very high, and for other industries it is very low. *Bécate* is a voluntary program, and so the firms that participate in the program must be firms with high learning rates. The authorities must consider this if the goal is to increase the number and types of firms participating in the program.

3.7 Final Remarks

The present study asks whether work experience in the informal sector can affect the career prospects of less-educated workers. The analysis focuses on two potential roles of informal sector jobs: accumulation of skills and screening of workers' ability. In the traditional queuing model of the informal sector with heterogenous workers' abilities, the hazard rate from the informal into the formal sector decreases with duration of informal sector employment. This study shows that, when informal sector jobs also enable workers to accumulate skills or employers to screen workers' abilities, the shape of the hazard function can be different from that predicted by the traditional queuing model. Human capital accumulation implies an increasing or U-shaped hazard due to the accumulation of skills (and the fact that more skilled workers leave the informal sector faster). Screening can generate a hump-shaped hazard if workers with observable ability leave (on average) faster than informal sector entrants, resulting in an increasing hazard; eventually, as more skilled workers leave faster, the hazard decreases with duration. These differences in the predicted hazard suggests a procedure to decide which role of informal sector jobs is more important.

The hazard function was estimated using an employment survey from Mexico. The estimated hazard reflects the hump-shaped pattern predicted by the screening model, indicating that informal sector jobs play an important role by screening young less-educated workers new to the labor market. Furthermore,

the estimation results reject the traditional queuing model with heterogenous workers' abilities, indicating that informal sector jobs provide some value above and beyond make-shift work while waiting to find a formal sector job.

Notice that this conclusion could break down if there is a third sector to which discouraged workers move, i.e. nonparticipation. If discouraged informal sector workers move to nonparticipation, and those workers moving out are mainly L-skilled workers, then the proportion of H-skilled workers would increase, pushing the average hazard function up; since H-skilled leave the informal sector at a faster rate, the proportion of H-skilled would decrease, pulling the average hazard function down. This alternate mechanism would produce a similar hump-shape pattern without employer learning. The question is whether young workers entering the labor market are discouraged and move to nonparticipation during the first years of their careers, which depends on the outside options of these workers. These outside options could be limited in developing countries like Mexico.

The employment survey used in this study is a rotating panel with a periodic follow-up, and so a significant fraction of the duration measures are interval-censored. In addition, for a good share of the spells the starting time is only known to fall within a twelve-month interval. These features of the data required the application of techniques for interval-censored failure time data, and a Monte Carlo study to investigate several alternatives for overcoming coarseness of the starting time of job spells. The latter is presented in chapter

4.

The parameters characterizing the employer learning process were inferred to determine how fast employers learn about their workers' abilities. The exercise suggests that employers learn about their workers' abilities at a much slower rate than that required by a government employment program, *Bécate*. This finding highlights the importance of a firm's involvement in the recruitment of workers participating in the program. In this way firms can minimize expected firing costs by recruiting candidates with a good match quality. Firm participation in the selection of candidates is allowed in the current format of *Bécate*.

3.8 Bibliography

ALBRECHT, J., L. NAVARRO, AND S. VROMAN (2006): "The Effects of Labor Market Policies in an Economy with an Informal Sector," Discussion Paper IZA DP No. 2141, The Institute for the Study of Labor (IZA).

——— (2009): "The Effects of Labour Market Policies in an Economy with an Informal Sector," *The Economic Journal*, 119(539), 1105–1129.

AMARAL, P. S., AND E. QUINTIN (2006): "A competitive model of the informal sector," *Journal of Monetary Economics*, 53, 1541–1553.

ARIAS, O., AND M. KHAMIS (2008): "Comparative Advantage, Segmentation and Informal Earnings: A Marginal Treatment Effects Approach," Discussion Paper IZA DP No. 3916, The Institute for the Study of Labor (IZA).

ARIAS, O., AND W. F. MALONEY (2007): "The *Razón de Ser* of the Informal Worker," in *Informality: Exit and Exclusion*, chap. 2, pp. 43–78. The World Bank, Washington, D.C.

BARRON, J. M., M. C. BERGER, AND D. A. BLACK (1997): *On-the-Job Training*. Kalamazoo, Michigan: Upjohn Institute for Employment Research.

BOSCH, M. (2006): "Job Creation and Job Destruction in the Presence of Informal Labour Markets," Working Paper, London School of Economics.

BOSCH, M., E. GONI, AND W. F. MALONEY (2007): "The Determinants of Rising Informality in Brazil: Evidence from Gross Worker Flows," Working Paper 4375, The World Bank.

CHAMBERLAIN, G. (1980): "Comment on 'The Analysis of Re-Employment Probabilities for the Unemployed' by T. Lancaster and S. Nickell," *Journal of the Royal Statistical Society. Series A (General)*, 143(2), 160.

CONGRESS (1970): *Federal Labor Law*. Mexico, D.F.

DELAJARA, M., S. FREIJE, AND I. SOLOAGA (2006): "An Evaluation of Training for the Unemployed in Mexico," Working Paper OVE/WP-09/06, Inter-American Development Bank.

DOLADO, J. J., M. JANSEN, AND J. F. JIMENO (2005): "Dual Employment Protection Legislation: A Framework for Analysis," Discussion Paper IZA DP No. 1564, The Institute for the Study of Labor (IZA).

FAJNZYLBER, P. (2007): "Informality, Productivity and the Firm," in *Informality: Exit and Exclusion*, chap. 6, pp. 157–178. The World Bank, Washington, D.C.

FARBER, H. S., AND R. GIBBONS (1996): "Learning and Wage Dynamics," *Quarterly Journal of Economics*, 111(4), 1007–1047.

FARRELL, D. (2004): "The hidden dangers of the informal economy," *The McKinsey Quarterly*, (3), 27–37.

FINKELSTEIN, D. M. (1986): "A Proportional Hazards Model for Interval-Censored Failure Time Data," *Biometrics*, 42(4), 845–854.

HAN, A., AND J. A. HAUSMAN (1990): "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics*, 5(1), 1–28.

HECKMAN, J. J., R. ROBB, AND J. R. WALKER (1990): "Testing the Mixture of Exponentials Hypothesis and Estimating the Mixing Distribution by

the Methods of Moments," *Journal of the American Statistical Association*, 85(410), 582–589.

HEMMER, H.-R., AND C. MANNEL (1989): "On the Economic Analysis of the Urban Informal Sector," *World Development*, 17(10), 1543–1552.

INEGI (2005): *Encuesta Nacional de Ocupación y Empleo 2005. Una Nueva Encuesta para México. ENOE*. Instituto Nacional de Estadística, Geografía e Informática, Aguascalientes, http://www.inegi.org.mx.

——— (2007): *Conociendo la base de datos de la ENOE*. Instituto Nacional de Estadística, Geografía e Informática, Aguascalientes, http://www.inegi.org.mx.

KALBFLEISCH, J. D., AND R. L. PRENTICE (1980): *The statistical analysis of failure time data*. Wiley, New York.

KIEFER, N. M. (1988): "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 26(2), 646–679.

LANCASTER, T. (1990): *The econometric analysis of transition data*. Cambridge University Press, New York.

LEVENSON, A. R., AND W. F. MALONEY (1998): "The informal sector, firm dynamics, and institutional participation," Working Paper 1988, The World Bank.

LEVY, S. (2007): "Can Social Programs Reduce Productivity and Growth? A Hypothesis for Mexico," Mimeo.

LOAYZA, N. V. (1996): "The economics of the informal sector: a simple model and some empirical evidence from Latin America," *Carnegie-Rochester Conference Series on Public Policy*, 45, 129–162.

MAGNAC, T. (1991): "Segmented or Competitive Labor Markets," *Econometrica*, 59(1), 165–187.

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

MEYER, B. D. (1990): "Unemployment Insurance and Unemployment Spells," *Econometrica*, 58(4), 757–782.

MORTENSEN, D. T., AND C. A. PISSARIDES (2003): "Taxes, subsidies and equilibrium labor market outcomes," in *Designing inclusion: tools to raise low-end pay and employment in private enterprise*, ed. by E. S. Phelps, pp. 44 – 73. Cambridge University Press, Cambridge; New York.

PRENTICE, R. L., AND L. A. GLOECKLER (1978): "Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data," *Biometrics*, 34(1), 57–67.

REBIÈRE, T. (2008): "Young workers' professional experience and access to high-skill jobs," Working paper, CERENE, University of Le Havre, France.

SCHNEIDER, F., AND D. H. ENSTE (2000): "Shadow Economies: Size, Causes, and Consequences," *Journal of Economic Literature*, 38(1), 77–114.

SUN, J. (2006): *The statistical analysis of interval-censored failure time data*. Springer, New York.

WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Mass.

WORLD BANK (2007): *World Development Report 2007: Development and the Next Generation*. The World Bank, Washington, DC.

Table 3.1: Summary Statistics by Education Group

Variable	Years of Education [0 , 6)	[6 , 9)	[9 , 12)
Age	21.18	20.48	20.75
Married	0.18	0.16	0.15
Monthly Earnings†	3385.70	3578.19	3424.24
Minimum Wage‡			
Zone A	0.09	0.11	0.14
Zone B	0.12	0.15	0.15
Zone C	0.79	0.74	0.71
Firm Size			
1-5	0.64	0.61	0.60
6-20	0.25	0.27	0.25
21+	0.11	0.13	0.15
Industry			
Construction	0.45	0.33	0.23
Manufacturing	0.21	0.21	0.20
Commerce	0.12	0.15	0.21
Services	0.22	0.31	0.36
Family Head Status§			
Formal Sector Job	0.10	0.16	0.22
Self-employed	0.13	0.11	0.12
Unemployed	0.02	0.01	0.02
Entrepreneur	0.06	0.07	0.09
Out of Labor Force	0.08	0.10	0.08
Number of Obs.	304	1,415	3,113

†Average monthly earnings in Mexican Pesos as of the 2nd half of December 2010. ‡Minimum wage by zone: A > B > C. §Employment status of the family head, when the family head is different from the individual in the sample.

Table 3.2: Distribution of Duration Data in the Sample (Number of Observations)

	Type of Sample			
Type of Interval	Flow	Stock 1†	Stock 2‡	Total
1-month	0	134	0	134
2-month	2	13	0	15
3-month	91	679	0	770
4-month	0	5	0	5
5-month	23	0	0	23
6-month	670	0	0	670
7-month	10	0	0	10
12-month	0	0	19	19
14-month	0	0	8	8
15-month	0	0	257	257
16-month	0	0	1	1
Right-censored	1,199	1,284	566	3,049
Total	1,995	2,115	851	4,961

†Workers with job start in the current or previous calendar year. ‡Workers with job start before the previous calendar year.

Table 3.3: Censoring in the Sample (Number of Observations)

	Type of Sample			
	Flow	Stock 1†	Stock 2‡	Total
Uncensored	796	831	285	1,912
Unemployed	224	314	57	595
Another risk§	412	279	185	876
Still working in IS	563	691	324	1,578
Total	1,995	2,115	851	4,961

§ Mainly composed by self-employment, but also includes unpaid family work, entrepreneurship, and out of the labor force. †Workers with job start in the current or previous calendar year. ‡Workers with job start before the previous calendar year.

Table 3.4: Summary Statistics of Duration Data in Weeks

	Years of Education		
	[0 , 6)	[6 , 9)	[9 , 12)
Complete Duration			
Mean	15.3	14.2	13.9
25th pctile	3.0	3.5	3.0
50th pctile	7.0	7.5	6.0
75th pctile	15.0	16.0	15.5
Elapsed Duration			
Mean	16.8	16.0	16.3
25th pctile	2.0	2.0	2.0
50th pctile	5.0	6.0	6.0
75th pctile	23.0	22.0	22.0

Note: For the purposes of getting these summary statistics, we imputed the interval-censored duration data using the midpoint in the interval.

Table 3.5: Estimated Piecewise Constant Hazard

	Years or Education		
	[0 , 12)	[0 , 9)	[9 , 12)
Firm size 6-20	0.4315	0.4390	0.4297
	(0.0553)	(0.0978)	(0.0675)
Firm size 21+	0.7777	0.9284	0.7225
	(0.0622)	(0.1127)	(0.0751)
Commerce Ind	0.2740	0.1316	0.3170
	(0.0740)	(0.1429)	(0.0875)
Services Ind	0.0139	0.1722	-0.0353
	(0.0653)	(0.1176)	(0.0790)
Construction Ind	0.0818	0.0792	0.0971
	(0.0706)	(0.1187)	(0.0883)
Graduate Grade 6	0.0902	0.0745	
	(0.1066)	(0.1076)	
Graduate Grade 9	0.2915		
	(0.0544)		
Married	0.2023	0.2942	0.1529
	(0.0633)	(0.1044)	(0.0798)
Family Head FS	0.2426	0.3403	0.2078
	(0.0554)	(0.1058)	(0.0650)
λ_1	0.0223	0.0270	0.0296
	(0.0040)	(0.0075)	(0.0055)
λ_2	0.1915	0.1866	0.5241
	(0.0293)	(0.0434)	(0.0916)
λ_3	0.0671	0.0495	0.1331
	(0.0148)	(0.0149)	(0.0216)
λ_4	0.0379	0.0415	0.0564
	(0.0057)	(0.0096)	(0.0084)
λ_5	0.0387	0.0405	0.0543
	(0.0055)	(0.0085)	(0.0069)
λ_6	0.0320	0.0364	0.0443
	(0.0041)	(0.0062)	(0.0048)
Log likelihood	-3,838.71	-1,344.01	-2,478.97
Number of Obs.	4,961	1,825	3,136

The omitted industry is Manufactures and the omitted firm size is 1-5 employees. The covariates also include: (i) a variable summarizing the number of self-employment scholarships approved in the state of residence relative to the size of the state's labor market, (ii) three dummies for the year of start of the IS-Job (1997-2004, 2005-2006, 2007-2008, 2009-2010), the first category is omitted, and (iii) a dummy for adjusted duration measures. Standard errors in parenthesis.

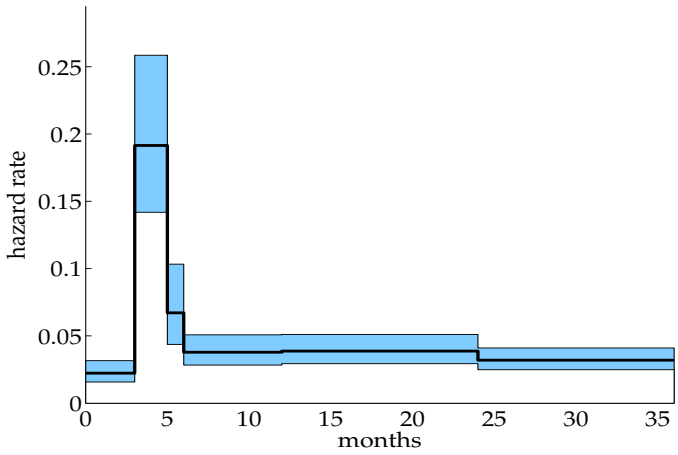
Table 3.6: Estimated Weibull and Log-logistic Hazards

	Weibull			Log-Logistic		
	Years or Education			Years or Education		
	[0 , 12)	[0 , 9)	[9 , 12)	[0 , 12)	[0 , 9)	[9 , 12)
Firm size 6-20	0.4347	0.4725	0.4170	0.6978	0.6481	0.7223
	(0.0549)	(0.0974)	(0.0667)	(0.0953)	(0.1601)	(0.1195)
Firm size 21+	0.8160	0.9662	0.7661	1.1837	1.1744	1.1798
	(0.0615)	(0.1121)	(0.0739)	(0.1135)	(0.1912)	(0.1417)
Commerce Ind	0.2875	0.1530	0.3265	0.4092	0.1542	0.5446
	(0.0734)	(0.1422)	(0.0865)	(0.1313)	(0.2385)	(0.1592)
Services Ind	0.0133	0.2019	-0.0492	-0.0006	0.0940	-0.0353
	(0.0649)	(0.1172)	(0.0783)	(0.1150)	(0.1941)	(0.1432)
Construction Ind	0.0860	0.1043	0.0963	0.1795	-0.0372	0.3344
	(0.0700)	(0.1185)	(0.0871)	(0.1206)	(0.1910)	(0.1569)
Graduate Grade 6	0.0563	0.0418		0.1639	0.1060	
	(0.1063)	(0.1073)		(0.1684)	(0.1716)	
Graduate Grade 9	0.2826			0.5403		
	(0.0541)			(0.0906)		
Married	0.2153	0.3066	0.1624	0.4348	0.5557	0.3441
	(0.0627)	(0.1039)	(0.0789)	(0.1112)	(0.1781)	(0.1427)
Family Head FS	0.2676	0.3529	0.2380	0.3665	0.4953	0.3167
	(0.0549)	(0.1053)	(0.0642)	(0.0975)	(0.1793)	(0.1166)
α	0.8630	0.8865	0.8539	1.6445	1.5083	1.7236
	(0.0247)	(0.0454)	(0.0294)	(0.0451)	(0.0736)	(0.0573)
T^*				5.73	5.96	5.49
Log likelihood	-4,022.24	-1,385.59	-2,626.72	-4,011.19	-1,401.62	-2,598.79
Number of Obs.	4,961	1,825	3,136	4,961	1,825	3,136

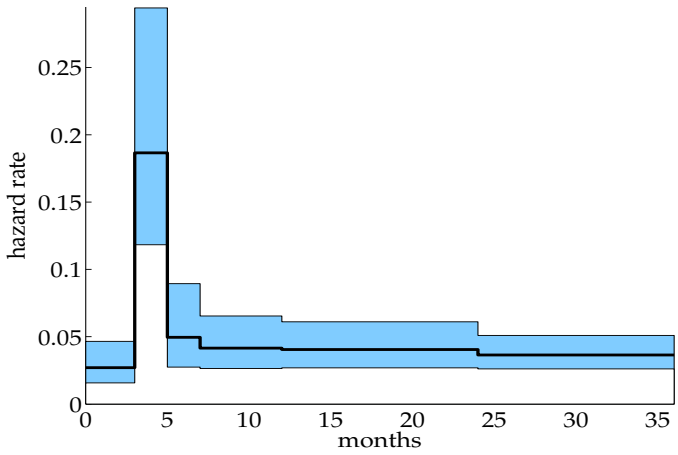
The omitted industry is Manufactures and the omitted firm size is 1-5 employees. The covariates also include: (i) a variable summarizing the number of self-employment scholarships approved in the state of residence relative to the size of the state's labor market, (ii) three dummies for the year of start of the IS-Job (1997-2004, 2005-2006, 2007-2008, 2009-2010), the first category is omitted, and (iii) a dummy for adjusted duration measures. T^* was computed using $x = \bar{x}$. Standard errors in parenthesis.

Figure 3.5: Piecewise Constant Baseline Hazard with 95% Pointwise Confidence Interval

(a) Years of Education: [0,12)



(b) Years of Education: [0,9)



(c) Years of Education: [9,12)

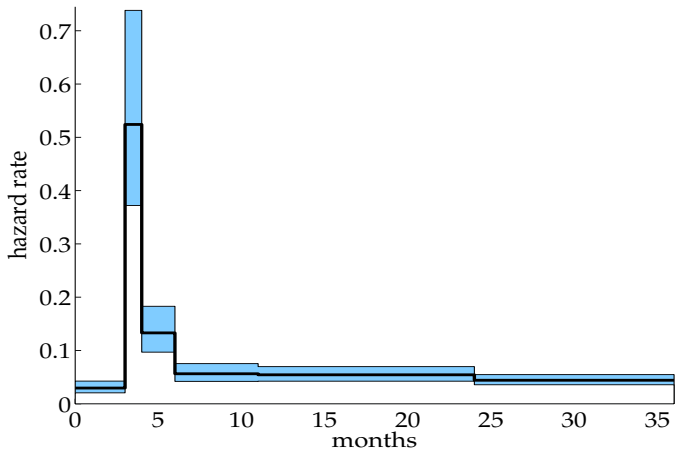
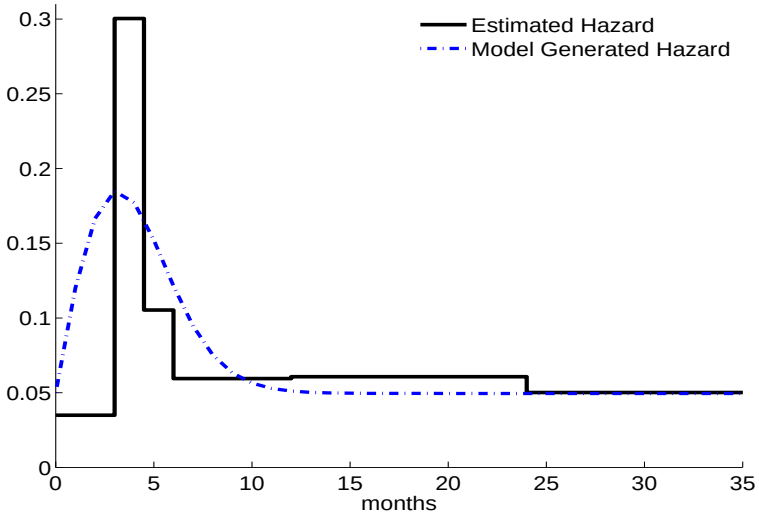


Figure 3.6: Estimated and Model-Generated Hazards



NOTE: The model-generated hazard uses $\bar{\mu} = 0.05$, $\mu(p_L) = 0.0495$, $\mu(p_H) = 1.0$, $\phi = 0.4833$, and $\sigma = 0.1478$. The estimated hazard uses $(\hat{\lambda}_1, \ldots, \hat{\lambda}_6)$ from the first column Table 3.5 and $\exp(\bar{x}'\hat{\rho}) = 1.57$.

Chapter 4

Stock Sampling with Interval-Censored Elapsed Duration: A Monte Carlo Analysis

4.1 Introduction

The length of time that individuals spend in a certain state, for example employment, is often at the center of applied studies in economics. Broadly speaking, there are two ways of obtaining duration data for this sort of study. One way is to sample individuals who enter the state of interest at some point during a fixed period of time. This sampling scheme is known as *flow sampling* (Wooldridge, 2002). Alternatively, the researcher can sample individuals who are already in the state of interest at a certain point in time. This sampling scheme is known as *stock sampling* (Wooldridge, 2002). Both sampling schemes are valid for inference, but the researcher must account for the type of sampling

in the estimation.

Ultimately, the researcher will use the sampling scheme that is more suitable for the study, and the one that is available in the data. Depending on the state of interest, flow sampling may require the fixed period of time to be long enough in order to observe a sufficient number completed spells.[1] In such a case, stock sampling may be more appropriate, since it only requires the researcher to follow the individual for a fraction of his or her spell in order to know the complete duration of the spell, given that the researcher knows the starting point of the spell.

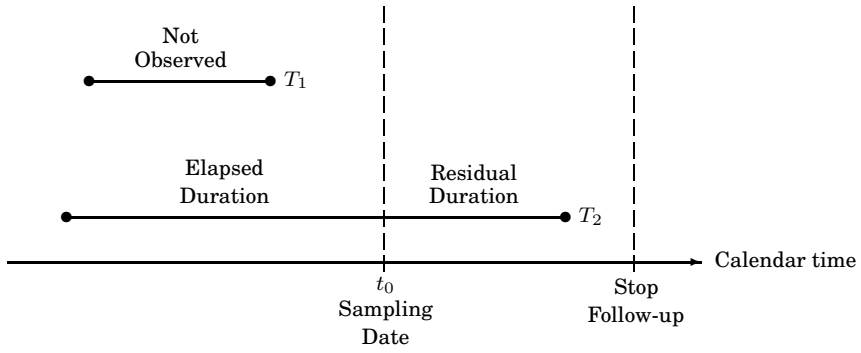
This paper focuses on a stock sampling scheme in which individuals who are already in a particular state are sampled at a given point in time, say t_0. At the sampling date, the length of time spent in the state of interest up to t_0 is recorded. This measure is called *elapsed duration*. After the sampling date, individuals are followed for a fixed period of time, and the length of time spent in the state after t_0 is recorded. This measure is called *residual duration*. Kalbfleisch and Prentice (1980) refer to this sampling scheme as *delayed entry*; Lancaster (1990) refers to it as *observation over a fixed interval* (see chap. 8, sect. 3.1); and Wooldridge (2002) refers to it as *stock sampling*. The current study follows the convention of calling it stock sampling.[2]

[1] An observed spell is said to be complete if both the beginning and the end of the spell are observed. Spells in which only the beginning is observed, but not the end, are called right-censored spells. Spells in which only the end is observed, but not the beginning, are called left-censored spells.

[2] Lancaster (1990) and Murphy (1996) call stock sampling a scheme in which only the elapsed duration of the individuals in the sample is observed, but there is no follow up of

Figure 4.1 describes the aforementioned sampling scheme. Note that short spells that started and ended before the sampling date, such as T_1, are not sampled, whereas, for sampled spells, such as T_2, the complete duration is the sum of the elapsed and the residual duration. Figure 4.1 illustrates a well known feature of stock sampling, that it produces a truncated sample because some spells are not observed. This is typically referred to as *left truncation* (Kalbfleisch and Prentice, 1980; Wooldridge, 2002), and the bias generated as *length-biased sampling*, because long spells are more likely to be sampled than short spells (Kiefer, 1988; Lancaster, 1990).

Figure 4.1: Stock Sampling



In order to account for left truncation in a stock sample, the likelihood function must incorporate the fact that long spells are sampled systematically more often. Let T be a random variable with density $f(t|x)$, representing the duration of the event of interest, where x is a set of time-invariant covariates, and let $\tilde{T}$ be a spell in the stock sample. Similarly, let s be the starting time of the event of interest and t_0 the stock sampling date, so that the elapsed duration

the individual after the sampling date (see Lancaster, 1990, chap. 8, sect. 3.3).

is $e = t_0 - s$. Then, in the stock sample, T is only observed if $T > e$. In other words, $\tilde{T} = T \mid T > e$, and so the density of $\tilde{T}$ is given by:

$$g(\tilde{t}|x) = f(t|x, T > e) = \frac{f(t|x)}{\Pr\{T \geq e|x\}}. \tag{4.1}$$

It is evident from (4.1) that knowledge of the starting time of the event is crucial for accounting for left truncation in a stock sample. This result is well known in the literature, e.g. Wooldridge (2002), Klein and Moeschberger (1997), Lancaster (1990), Kalbfleisch and Prentice (1980).

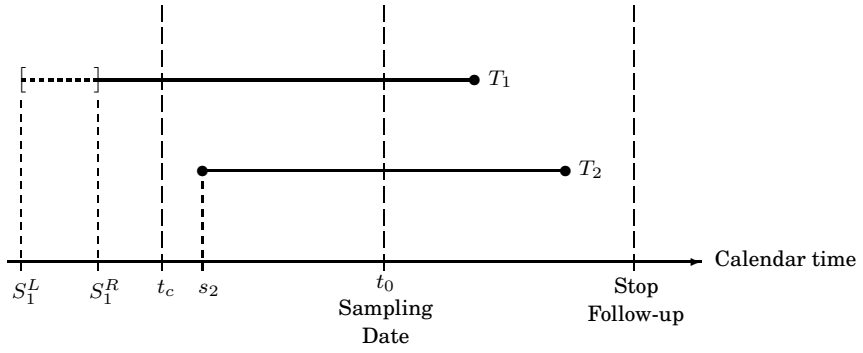
This paper addresses a further complication in the stock sampling scheme that has not been addressed in the literature and that arises in some socioeconomic surveys. The stock sampling scheme in this paper is identical to the one described above with the difference that, for some spells, the starting time is only known to be contained within some interval. Let t_0 be the stock sampling date, and let $t_c < t_0$ be some date such that, if $t_c < s < t_0$, then s is observed, but if $s < t_c$, then, only an interval $[S^L, S^R]$ containing s is observed. Equivalently, for spells that started after t_c, the elapsed duration e is observed, but for spells that started before t_c, it is only known that e is contained in the interval $[E^L, E^R]$.[3]

Figure 4.2 describes the sampling scheme that this paper explores. For spells that started before t_c, such as T_1, the starting time is only known to be contained in an interval (interval $[S_1^L, S_1^R]$ in the picture), whereas, for spells

[3]Note that this problem is not as severe as left-censoring, in which the starting time, and hence the elapsed duration, is not known at all.

that started after t_c, such as T_2, the exact starting time is observed (s_2 in the picture). Note that the density of observed spells in this sample is different from (4.1) because for some spells, e is only known to be contained in the interval $[E^L, E^R]$.

Figure 4.2: Stock Sampling with Interval-Censored Starting Time



This sampling scheme is common when obtaining job duration data from surveys that are implemented as rotating panels. Two examples are the National Survey of Occupation and Employment from Mexico, ENOE (its acronym in Spanish), and the Monthly Employment Survey from Brazil, PME (its acronym in Portuguese). Both surveys have been used in studies of duration of employment in Mexico (see Chapter 3) and in Brazil (Ulyssea and Szerman, 2006). The PME provides the exact elapsed duration if the respondent's current job started within the last two years, but only the number of years elapsed if the job started more than two years before the interview. In the ENOE, the starting time of the job is known if the respondent's current job started during the previous calendar year, but only the starting year is known if the job started

before the previous calendar year.

Table 4.1 shows that in both surveys the fraction of spells with interval-censored elapsed duration or interval-censored starting times is significant. In the ENOE, 72% of paid employees in the first quarter of 2010 started their current jobs before the previous calendar year, and so the survey only records the year when such a job started. In the PME, 60% of the paid employees in January of 2012 started their current jobs more than two years before the interview, and so the survey only records the number of years elapsed since the job started.

Table 4.1: Elapsed Duration in the ENOE and PME

Elapsed Duration	ENOE		PME	
	Num. Obs.	%	Num. Obs.	%
Exact	20,499	27.37	12,875	39.51
Interval-Censored	54,399	72.63	19,713	60.49
Total	74,898		32,588	

Source: INEGI for ENOE, IBGE for PME. Data from the ENOE is for the first quarter of 2010. Data from PME is for January of 2012. The table only includes paid employees.

As Table 4.1 suggests, spells with interval-censored elapsed duration cannot be ignored. The goal of this study is to investigate different alternatives for overcoming this coarseness by imputing the interval-censored elapsed duration and performing a Monte Carlo analysis to gauge the properties of the estimators, focusing on the unbiasedness of the estimators. The Monte Carlo analysis is based on simulated duration data that resembles the sampling scheme of

the ENOE, which is further explained below. The interval-censored elapsed duration is imputed using: (1) the lower bound of the interval containing the elapsed duration, (2) the midpoint of the interval, and (3) the upper bound of the interval. The results indicate that using the midpoint to impute the interval-censored elapsed duration outperforms the alternatives.

The study is organized as follows. Section 4.2 describes the problem at hand and a direct approach to overcome the coarseness of starting times. It also describes the imputation methods as an alternative to the direct approach for estimation. Next, in Section 4.3, the simulation algorithm is presented. Section 4.4 presents the results from the Monte Carlo simulations and Section 4.5 presents an application of the imputation methods to duration data from the ENOE. Finally, Section 4.6 concludes.

4.2 Interval-Censored Starting Times

The problem of interval-censored starting times introduced in the previous section can be addressed as one would address the problem of left-censoring (see Wooldridge, 2002, exercise 20.8). Using the same notation as before, let T be a random variable with density $f(t|x;\theta)$ representing the duration of the event of interest, where x is a set of time-invariant covariates and θ is the vector of parameters of interest characterizing the duration model. Let t_0 be the stock sampling date, and S the starting time of the event with density $k(s|x;\eta)$, where

η is the vector of parameters characterizing the distribution of S. Also, as before, the elapsed duration is defined as $E = t_0 - S$, and the residual duration is defined as $U = T - t_0 + S$. Suppose that, conditional on the covariates x, starting times are independent of the duration variable. Then, the joint density of T and S is given by $g(t, s|x; \theta, \eta) = f(t|x; \theta)k(s|x; \eta)$. Then, using the change of variable technique, it is straightforward to show that the density of U is given by:

$$h(u|x; \theta, \eta) = \int_{S^L}^{S^R} f(u + t_0 - s|x; \theta)k(s|x; \eta)ds. \tag{4.2}$$

Next, suppose that, after t_0, individuals in the stock sample are only followed during a fixed interval of time C. Then, if $U > C$ the spell will be right-censored. Hence, the probability that the spell is right-censored is given by:

$$\Pr\{U > C|x\} = 1 - \int_0^C h(u|x; \theta, \eta)du = 1 - H(C|x; \theta, \eta). \tag{4.3}$$

Finally, to obtain the contribution to the likelihood function from a spell with an interval-censored starting time, it is necessary to account for stock sampling. For this, recall that a spell t in a stock sample is observed if and only if $t > t_0 - S$. Thus the probability that a spell with interval-censored starting time is sampled is given by:

$$\Pr\{T > t_0 - S|x\} = 1 - \int_{S^L}^{S^R} \int_0^{t_0 - z} g(t, s|x; \theta, \eta)dtds = \int_{S^L}^{S^R} \Big[1 - F(t_0 - s|x; \theta)\Big] k(s|x; \eta)ds. \tag{4.4}$$

Hence, the contribution to the likelihood from a spell with residual duration u_i,

and interval-censored starting time $[S_i^L, S_i^R]$, is given by:

$$L_i(\theta, \eta | x_i) = \frac{h(u_i|x_i;\theta,\eta)^{d_i}\left[1 - H(C|x_i;\theta,\eta)\right]^{(1-d_i)}}{\int_{S_i^L}^{S_i^R}\Big[1 - F(t_0 - s|x_i;\theta)\Big]k(s|x_i;\eta)ds}, \tag{4.5}$$

where d_i, is an indicator equal to 1 for completed spells and 0 for right-censored spells. Hence with knowledge of $k(\cdot|\cdot)$, it is possible to estimate the vector of parameters (θ, η). In many cases, the likelihood function will involve integrals for the contribution of spells with interval-censored starting times. For example, in the widely used case where T follows a Weibull distribution, and assuming that S follows a uniform distribution in $[S^L, S^R]$, the cumulative distribution function of U at the right-censoring time is given by:

$$H(C|x;\alpha,\beta) = \frac{1}{S^R - S^L}\left(\int_0^C e^{-\exp(x'\beta)(u+t_0-S^R)^\alpha}du - \int_0^C e^{-\exp(x'\beta)(u+t_0-S^L)^\alpha}du\right).$$

This has no closed-form solution. In such cases, the researcher can apply techniques of numerical integration and proceed with the estimation of (θ, η). The following section describes an alternative procedure to estimate θ by imputing interval-censored starting times.

4.2.1 Alternative for Estimation: Imputed Starting Times

The previous section develops a methodology to estimate the parameter of the duration model, θ. The methodology depends on the assumption that starting times are independent of the duration variable, requires knowledge of the

distribution of starting times, and involves no-closed form expressions.[4] This section explores an alternative for estimation of θ by imputing a missing starting time using the interval which contains the starting time. Notice that this method is equivalent to imputing the missing elapsed duration using the interval that contains the elapsed duration. In what follows, the imputation methods are discussed in terms of the elapsed duration.

Three imputation methods are explored, using: (1) the lower bound, (2) the upper bound, and (3) the midpoint of the interval $[E^L, E^R]$. Let $\hat{E}$ be the imputed elapsed duration, then, the imputation methods are:

1. $\hat{E} = E^L$
2. $\hat{E} = \dfrac{E^L + E^R}{2}$
3. $\hat{E} = E^R$

A fourth imputation method, which consisted in replacing the missing elapsed duration with a random draw from the interval $[E^L, E^R]$, was also considered. The random draw was based on a uniform distribution on $[E^L, E^R]$. This imputation method produced results very similar to the results using the midpoint of the interval. Hence, only the results using the midpoint of the interval are presented here.

[4] Alternatively, the assumption on independence could be relaxed and the distribution of starting times could be nonparametrically estimated. This will make estimation even more computationally demanding.

Finally, the contribution to the likelihood of a spell with interval-censored starting time is given by:

$$L_i(\theta|x_i) = \frac{f(\hat{t}_i|x_i;\theta)^{d_i}\left[1 - F(\hat{e}_i + C|x_i;\theta)\right]^{(1-d_i)}}{1 - F(\hat{e}_i|x_i;\theta)} \tag{4.6}$$

where d_i is the indicator defined before, $\hat{e}_i$ is the imputed elapsed duration, $\hat{t}_i = u + \hat{e}_i$ is the imputed duration, and $F(\cdot|\cdot)$ is the cumulative distribution of T. The parameters from the duration model, θ, can be estimated by maximum likelihood.

The question is when do we want to impute the elapsed duration instead of using the exact likelihood to estimate the parameters of the duration model. The choice will depend on the problem at hand. The main advantage of imputations is the savings on computation time, which could be expensive in some set ups. For example, in the case of continuous duration presented in the previous section, the exact likelihood involves integrals with no closed form, in which case the researcher will have to use Monte Carlo integration or quadratures. Even though this is feasible, it could be computationally expensive. Similarly, in the case of discrete duration, evaluating the exact likelihood could be computationally expensive if the duration variable has a non-trivial portion of observations with a fine measure of duration, e.g. weeks, and so the researcher would like to use the finest possible measure. That is, suppose that some of the duration measures are known up to the week, and for some durations with interval-censored starting times the duration is only known to be contained in a 52-week interval. In that case, we know that the true duration could be one

of the 52 possible durations, which implies computing the likelihood 52 times (because it is the union of mutually exclusive events: week 1, week 2, . . ., or week 52). Again, this could be computationally expensive. The goal of this paper is to explore the properties of the estimators when using imputed measures in the likelihood instead of the the exact likelihood.

The present paper performs a Monte Carlo analysis to explore the finite-sample performance of the estimator under each of the three imputation methods. The simulated duration data resembles the duration data obtained from surveys like the ENOE or PME. The simulation algorithm is described in the following section.

4.3 Simulation of Survey Data

This section explains how the duration data are simulated. Since the simulation exercise is done to explore different alternatives to address the problems faced when working with duration data obtained from surveys such as the ENOE or the PME, the simulation is tailored to match the features of duration data obtained from these surveys. In particular, this paper focuses on employment duration data obtained from the ENOE. The simulation algorithm to obtain the continuous-time duration data is explained in the next section.

4.3.1 Simulation Algorithm

A continuous-time data set is generated first. This data set is used later to generate different data sets with interval-censored starting times. The continuous-time data are obtained assuming that T follows a Weibull-Gamma distribution, with a hazard function given by:

$$\lambda(t|x,\nu) = \mu\alpha t^{\alpha-1}\nu, \tag{4.7}$$

where α is the measure of duration dependence; μ is a scale parameter, which is parameterized as $\mu = \exp\{\beta_0 + \beta_1 x\}$ to account for observed heterogeneity; and the parameter ν represents unobserved heterogeneity, which is assumed to be distributed Gamma(κ, δ) with density:

$$g(\nu) = \frac{1}{\Gamma(\kappa)}\delta^{\kappa}\nu^{\kappa-1}e^{-\delta\nu},$$

so that $\mathrm{E}[\nu] = \kappa/\delta$ and $\mathrm{Var}[\nu] = \kappa/\delta^2$. The cumulative distribution function of T conditional on ν is given by:

$$F(t|x,\nu) = 1 - \exp\{-\mu t^{\alpha}\nu\}. \tag{4.8}$$

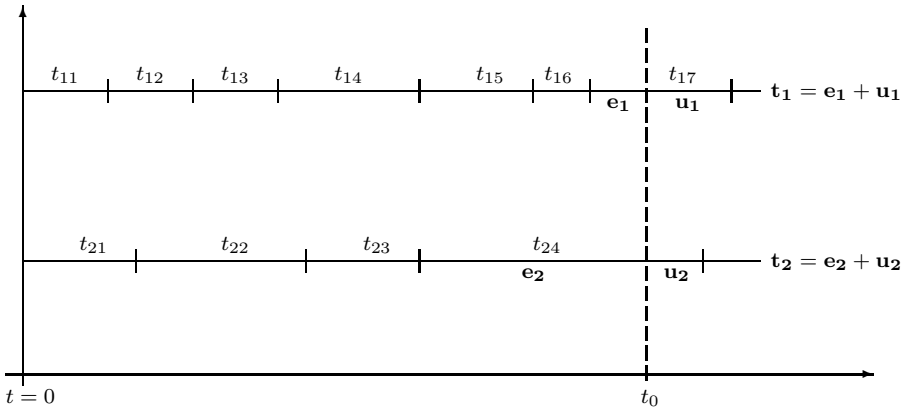
The distribution of ν is normalized by setting $\mathrm{E}[\nu] = 1$, which implies setting $\kappa = \delta$, consequently $\mathrm{Var}[\nu] = 1/\delta$. As a result, small values of δ imply that a large portion of the variation in the duration variable is due to unobserved heterogeneity, whereas large values of δ imply that unobserved heterogeneity is only responsible for a small fraction of the variation in the duration variable. When δ grows indefinitely, the distribution of T converges to a Weibull distribution without unobserved heterogeneity (see Cameron and Trivedi, 2005).

In the literature, duration data is usually simulated assuming that that these data arise from a flow sampling scheme (e.g. Ridder, 1987; Heckman and Singer, 1984; Baker and Melino, 2000), but this paper is mainly interested in a stock sampling scheme. To generate stock sampling data the duration data are simulated in the same way as a *renewal process*. The description of the renewal process presented below is based on the description provided by Lancaster (1990).

Refer to Figure 4.3. Let the subindex i identify a sequence of generations of individuals with the same observable and unobservable characteristics, and let the subindex j identify a particular member of a given sequence, so that the duration t_{ij} refers to the duration of member j from the sequence of generations of individuals i. Suppose that a population of workers starts employment at some time $t = 0$. Consider a sequence of generations of individuals with observable and unobservable characteristics (x_i, ν_i). Each member of the population in the first generation is employed for a random period of time T_{i1} with distribution $F(t|x_i, \nu_i)$ and realization t_{i1}. When a member of the first generation exits the state of employment, this member is replaced with another member (the second generation) who is employed for a random period of time T_{i2} with distribution $F(t|x_i, \nu_i)$ and realization t_{i2}. This process repeats indefinitely. Stock sampling takes place at some time $t = t_0$. In Figure 4.3, two hypothetical generation sequences of the simulated population are sampled. At the stock sampling date, t_0, the first sequence is in its seventh generation, while the second sequence is

in its fourth. At the stock sampling date, the elapsed duration of each member is $\mathbf{e_1}$ and $\mathbf{e_2}$, and their residual duration is $\mathbf{u_1}$ and $\mathbf{u_2}$, and so the sampled employment spells are $\mathbf{t_1 = e_1 + u_1}$ and $\mathbf{t_2 = e_2 + u_2}$.

Figure 4.3: Simulation as a Renewal Process



For each sequence of generations i, the stock sample data generation uses the following steps:

1. Draw x_i from a $\mathrm{N}(\mu_x, \sigma_x^2)$, to obtain $\mu_i = \exp\{\beta_0 + \beta_1 x_i\}$.

2. Draw ν_i from a $\mathrm{Gamma}(\delta, \delta)$ distribution.

3. Start with $j = 1$, compute the duration spell t_{ij} as follows:

   (a) Draw Y from a Uniform[0,1].

   (b) Compute t_{ij} using the inverse of the cumulative distribution function (4.8) as follows:

$$t_{ij} = F^{-1}(Y|x_i, \nu_i) = \left[-\frac{\ln(1-Y)}{\mu_i \nu_i}\right]^{1/\alpha}$$

4. Compute the cumulative duration for the sequence of spells up to generation j: $\overline{\overline{T}}_{ij} = \sum_{k=1}^{j} t_{ik}$.

5. If $\overline{\overline{T}}_{ij} > t_0$, then stop and go to 6, otherwise go back to 3, increase j by 1, and repeat process.

6. Once the stock sampling date is reached, compute the residual, the elapsed, and the complete duration, respectively, as:

$$u_i^* = \overline{\overline{T}}_{ij} - t_0$$

$$e_i^* = \overline{\overline{T}}_{ij} - u_i^*$$

$$t_i^* = e_i^* + u_i^*$$

This process is repeated for $i = 1, 2, \ldots, 3000$, that is the sample size is $N = 3,000$. Notice that the draw of the observed and unobserved heterogeneity components, x_i and ν_i respectively, is done only once for each sequence of generations of the population and stays constant during the repeated draws from the uniform distribution Y.

Six different sets of parameters of the data generating process were chosen. These parameter sets account for every combination of three cases of duration dependence with two cases of unobserved heterogeneity. The three cases of duration dependence are: (1) negative duration dependence, $\alpha < 1$; (2) no duration dependence, $\alpha = 1$; and (3) positive duration dependence, $\alpha > 1$. The two cases of unobserved heterogeneity are: (1) no unobserved heterogeneity, large value of δ; and (2) unobserved heterogeneity, small value of δ. The scale parameter, μ,

is chosen to match the observed stock-sample average duration in the informal sector from the ENOE in weeks, which is 82.12 weeks. Since $\mu = \exp\{\beta_0 + \beta_1 x\}$, the parameter used to match the survey data is β_0, while β_1 is set to $\beta_1 = 1$. The parameter β_0 is chosen using the simulated samples without unobserved heterogeneity. The same value of β_0 is used for the samples with unobserved heterogeneity, so that the only difference between two samples with the same degree of duration dependence is in the value of δ.

Only one covariate x is considered. The same covariate is used for all simulated samples and for all six parameter sets. This covariate is drawn from a $\mathrm{N}(\mu_x, \sigma_x^2)$. The mean of this distribution is set to $\mu_x = 0$, and its variance is set to $\sigma_x^2 = 0.25$. The choice of variance follows Baker and Melino (2000): σ_x^2 is chosen so that the R^2 from a regression of the simulated $\ln(T)$ on the simulated x is similar to the R^2 of a similar regression using the duration data and a set of covariates from the ENOE.[5]

The six parameter sets are presented in Table 4.2.

Generating Survey-like Samples

Once the continuous-time duration data have been generated, it is straightforward to generate samples with features similar to those of the ENOE. The first feature is right-censoring. For spells with residual duration $u_i > C$ the

[5]The duration data from the ENOE contains some interval-censored spells. In order to fit this regression, these spells are imputed as $\tilde{T}_i = L_i + u \cdot (R_i - L_i)$, where u is drawn from a uniform distribution in $[0, 1]$ and $(L_i, R_i]$ is the interval containing the actual duration.

Table 4.2: Parameters of Data Generating Process

Parameter Set	β_0	β_1	α	δ
1	-1.03	1	0.5	100
2	-1.03	1	0.5	1
3	-3.85	1	1.0	100
4	-3.85	1	1.0	1
5	-6.40	1	1.5	100
6	-6.40	1	1.5	1

residual duration is set equal to C, where C is a fixed right-censoring period (equal to 52 weeks in the case of the ENOE). The second feature, and the object of this paper, is interval-censored elapsed durations. For spells with starting times before the previous calendar year (i.e. 52 weeks before the stock sampling date), only the year when the spell started is known. For these spells, the elapsed duration is only known to be contained in a 52-week interval $[E_i^L, E_i^R]$.

4.4 Simulation Results

A total of 100 continuous-time data sets, each of size $N = 3,000$, were generated for each parameter set using the algorithm and parameters presented in the previous section. Using these continuous-time data sets two sample designs were generated. These are described in Table 4.3.

Table 4.5 presents the estimation results using sample design CONTA but ignoring stock sampling. That is, in the likelihood function (4.6), the term in the denominator $[1 - F(e_i^*|x_i;\theta)]$ is ignored.[6] The table presents the true

[6]Note that in this case the denominator is in terms of e_i^*, and not $\hat{e}_i$ because in this sample

Table 4.3: Sample Designs with Continuous-Time Data

Sample Design	Residual Duration	Elapsed Duration
CONTA	$u_i = \min\{u_i^*, 52\}$	$e_i = e_i^*$
CONTB	$u_i = \min\{u_i^*, 52\}$	If $e_i^* < 52 \Rightarrow e_i = e_i^*$ If $e_i^* > 52 \Rightarrow [E_i^L, E_i^R]$ 52-week interval

parameter values, the average, and the standard deviation of the point estimates using the simulated samples. It is evident from the table that ignoring stock sampling severely affects the estimates of the parameters of the duration model. The upward bias in α is linked to the downward bias in δ. The estimated parameters indicate a high degree of unobserved heterogeneity even when this feature is not present in the simulated data (recall that $\delta = 100$ implies a low degree of unobserved heterogeneity).

Table 4.6, on the other hand, presents the estimation results using sample design CONTA and accounting for stock sampling. All estimates, with the exception of the estimate of δ, include the true parameter value within one standard deviation, indicating that the estimator is unbiased. The large value of the estimate of δ for parameter sets 1, 3, and 5 results from the fact that the likelihood function is relatively flat with respect to δ, and an estimate of $\hat{\delta} = 100$ yields basically the same likelihood as an estimate of $\hat{\delta} = 10,000$. In such cases, the researcher may want to drop unobserved heterogeneity in the model.

the "exact" elapsed duration is observed.

From all six parameter sets, notice that, in the parameter set with negative duration dependence and unobserved heterogeneity (Parameter Set 2), the average estimate of α is: (i) the farthest away from the true α, and (ii) above the true α. This result was noted by Baker and Melino (2000) for the case of the non-parametric MLE. As they explain, it results from the fact that unobserved heterogeneity can produce negative duration dependence, even when the latter does not exist. Hence, when the optimization algorithm converges to a low estimate of δ, it converges to a large estimate of α, which is larger than the true parameter. In addition, note that the estimates of β_1 for this parameter set are the farthest away from the true parameter, but no more than a standard deviation away. The estimate for β_0 in this parameter set, however, is downward biased.

The results in Table 4.6 serve as a benchmark because the estimation accounts for stock sampling when the elapsed duration is known exactly. Table 4.7 presents the results from the estimation using sample design CONTB, in which the elapsed duration is only known within a 52-week interval if the elapsed duration is longer than a year. The estimation results suggest that the three imputation methods yield satisfactory results for this case. Notice that using $\hat{E} = E^L$ usually yields estimates larger than using $\hat{E} = (E^L + E^R)/2$, which in turn also yields estimates larger than using $\hat{E} = E^R$. However, almost all estimates are very close to the true parameter values. Thus, if all that is extraordinary in the data set is the interval-censoring of starting times, the

numbers in the table suggest that using either of the three imputation methods will yield good estimates of the duration dependence parameters and the coefficients on the covariates.

Notice that, for the Parameter Set 2, the estimates for β_0 are downward biased. However, this downward bias also occurs in the benchmark case in Table 4.6, which has an "ideal data set", and so interval-censored starting times cannot be held responsible for this problem.

4.5 Duration Data in the ENOE

4.5.1 Duration of Informal-Sector Employment in the ENOE

The ENOE is a rotating panel in which households are followed over 12 months with periodic visits every three months. Hence, the household is visited five times over the course of a year. This survey is widely used in studies of the informal sector as it provides the means to determine whether employed individuals belong to the formal or the informal sector (e.g. Chapter 3, Flores-Vazquez, 2011). The informal sector is composed of all individuals holding a job that does not comply with the labor regulations, e.g. social security, minimum wage, severance pay.

Given the relevance of the ENOE for the study of the informal sector, this

paper focuses on measures of employment duration in the informal sector before moving to the formal sector, that is, measures of the length of time that passes between the point in time when an individual starts an informal sector job and the point in time when such an individual gets a formal sector job.[7] From the information collected by the ENOE, it is possible to sample individuals who are already employed in the informal sector. Since the ENOE only follows individuals over 12 months, if an employment spell has not yet finished by the last interview, the spell is right-censored. In addition, if the employment spell started before the previous calendar year, only the year when employment started is observed.[8]

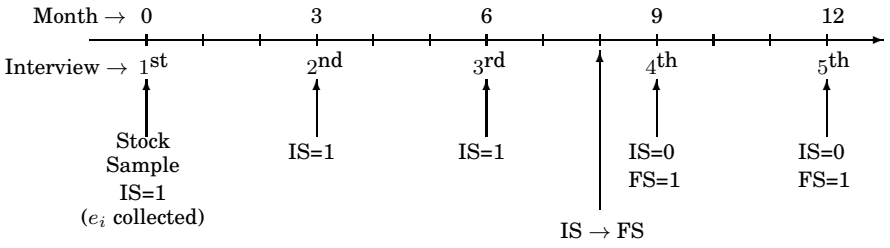
Figure 4.4 describes an example of an informal-sector employment spell constructed from the ENOE. In the first interview, the ENOE collects information about the starting time of the informal sector job. This information is used to construct the measure of elapsed duration, that is, the time elapsed from the job start to the first interview (e_i months in Figure 4.4). From the information collected in the following visits, it is possible to construct a measure of the residual duration, that is, the time elapsed from the first interview until the individual gets a formal sector job (8 months in Figure 4.4). The complete

[7] In practice, an individual currently employed in the informal sector can "move" to the formal sector within the same firm or with another firm. The data suggest that in the majority of the transitions from the informal into the formal sector the individual changes firms.

[8] The ENOE only collects information about the starting time of the current job in the Long Form of the ENOE which is answered at least once by each panel but not always in the first visit to the household. The simulation exercise in the current study assumes that the long form of the ENOE was answered in the first visit to the household.

duration from the stock sample is the sum of the elapsed and the residual duration ($e_i + 8$ months in Figure 4.4). Finally, notice that if the respondent is still employed in the informal sector by the fifth interview, the employment spell will be right-censored, in which case it is only possible to know that the informal-sector employment spell is larger than $e_i + 12$ months.

Figure 4.4: Stock Sampling from the ENOE



Note: IS is and indicator variable equal to 1 if the individual holds an informal sector job. IS = Informal Sector, FS = Formal Sector, e_i = Elapsed Duration of individual i.

Under the best of circumstances, the information of the ENOE for each respondent would include: (i) the exact number of months of elapsed duration, e_i, and (ii) the exact number of months of residual duration, u_i. However, this is not always the case. In many cases, the exact month when the respondent makes a transition from the informal into the formal sector is not observed. Using the example in Figure 4.4 again, it is known that, at the time of the third interview, the respondent was employed in an informal sector job, and that, by the time of the fourth interview, the respondent was employed in a formal sector job. However, in many cases, the exact month of transition from the

informal into the formal sector is not known.[9] In such cases, the respondent's residual duration is interval-censored. In terms of Figure 4.4, the residual duration is only known to be in the interval $[6, 9)$ months. Consequently, the complete duration in the informal sector is only known to be contained in the interval $[e_i + 6, e_i + 9)$ months, which includes the correct monthly measure of complete duration, $e_i + 8$ months.

4.5.2 Simulation Results

Three additional sample designs were generated to explore the properties of the estimators under each imputation method when using duration data like those provided by the ENOE. These additional sample designs are described in Table 4.4.

Table 4.4: Sample Designs with Interval-Censored Data

Sample	Residual Duration	Elapsed Duration
INTCA	$(L_i, R_i]$ 13-week interval	$e_i = e_i^*$
INTCB	$(L_i, R_i]$ 13-week interval	If $e_i^* < 52 \Rightarrow e_i = e_i^*$ If $e_i^* > 52 \Rightarrow [E_i^L, E_i^R]$ 52-week interval
INTCC	$(L_i, R_i]$ 13-week interval	If $e_i^* < 52 \Rightarrow [E_i^L, E_i^R]$ 4.3-week interval If $e_i^* > 52 \Rightarrow [E_i^L, E_i^R]$ 52-week interval

NOTE: Right-censored spells have residual duration in the interval $[52, \infty)$.

[9] This could result for many reasons. One of them is because the respondent made the transition within the same firm, and so there is no actual recollection of the time when the respondent started the formal sector job, because for the practical purposes it is the same job for the respondent. Another reason is due to the fact that the Long Form of the ENOE is not used in all five interviews, and the information about the start of a job is not available then.

Each of the sample designs in Table 4.4 have interval-censored residual duration, but differ in how the elapsed duration is observed. Sample design INTCC is the one that resembles the duration data from the ENOE more closely. Since the continuous-time duration data is generated to "mimic" weekly data, if all that is known is the number of months of elapsed duration, then it is known that the exact elapsed duration is contained in a 4.3-week interval. On the other hand, if all that is known is the number of years of elapsed duration, then it is known that the exact elapsed duration is contained in a 52-week interval. Samples designs INTCA and INTCB are provided to introduce each of these features one at a time and to be able to gauge the effect on the estimates of each of these features.

The estimation results using simulated sample from the INTCA design are presented in Table 4.8. The results in the table indicate that interval-censoring of residual duration does not affect the properties of the estimators. The estimation results are almost as good as the results when using the continuous-time data presented in Table 4.6. Next, Table 4.9 presents the results using simulated samples from the INTCB design. Once again, the estimation results using either of the three imputation methods are quite satisfactory. In most cases, the true parameter is not further than one standard deviation from the average point estimate. As is usually the case, in Parameter Set 2, the true parameters are more difficult to recover. A similar difficulty arose when using the continuous-time data presented in Table 4.7, hence interval-censored residual

duration is not responsible for this problem.

Finally, Table 4.10 presents the more realistic and interesting case, sample design INTCC, in which the elapsed duration is measured in months or years and the residual duration is interval censored. Sample design INTCC is the one that most closely resembles the duration data generated from the ENOE as described in section 4.5.1. For this sample, all measures of residual duration are only known to be contained in a 13-week interval (0,13], (13,26], (26,39], and (39,52], except for those that are right-censored, which are also interval-censored in the interval $(52, \infty)$. On the other hand, all starting times are also interval-censored. For spells that started during the previous calendar year, the starting time is only known to be contained in the 4.3-week interval. For spells that started before the previous calendar year, the starting time is only known to be contained in a 52-week interval. The estimation results, presented in Table 4.10, indicate that using the midpoint in the interval to impute the missing elapsed duration yields the best estimates of β_1 and α, which are usually the parameters in which the researcher is most interested. Once again, the estimates for β_0 are rather poor.

4.6 Final Remarks

This paper explores an alternative for estimation of a duration model when the data is obtained from a stock sample and the starting times of the spells

are only known to be contained in an interval. A direct approach for estimation in this case involves "integrating-out" the missing starting times. This approach requires assumptions about the distribution of starting times and, in most cases, involves non-closed forms in the likelihood function that would require the use of numerical integration. The alternative is to simply impute the missing starting times using the information provided by the interval that contains them. Since this paper is interested in the finite sample properties of the estimated parameters under each imputation method, it provides a Monte Carlo analysis to explore these properties. The property of interest is the unbiasedness of the estimator, particularly for the duration dependence parameter and the coefficients on the covariates.

The results indicate that, if the researcher has access to continuous-time duration data and interval-censored starting times are the only unusual feature of the data, then using either the lower bound, the upper bound, or the midpoint of the interval produces unbiased estimates. However, in the case which arises in commonly used surveys where the elapsed duration is measured in months or years, the results indicate that using the midpoint outperforms the alternatives.

4.7 Bibliography

BAKER, M., AND A. MELINO (2000): "Duration dependence and nonparametric heterogeneity: A Monte Carlo study," *Journal of Econometrics*, 96(2), 357 – 393.

CAMERON, A. C., AND P. K. TRIVEDI (2005): *Microeconometrics : Methods and Applications*. Cambridge University Press, Cambridge ; New York.

FLORES-VAZQUEZ, I. M. (2011): "Health Insurance Provision in Mexico: A Two-Sector Equilibrium Search Model," Working paper, , New York University.

HECKMAN, J., AND B. SINGER (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2), 271–320.

KALBFLEISCH, J. D., AND R. L. PRENTICE (1980): *The statistical analysis of failure time data*. Wiley, New York.

KIEFER, N. M. (1988): "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 26(2), 646–679.

KLEIN, J. P., AND M. L. MOESCHBERGER (1997): *Survival analysis : techniques for censored and truncated data*. Springer, New York.

LANCASTER, T. (1990): *The econometric analysis of transition data*. Cambridge University Press, New York.

MURPHY, A. (1996): "A piecewise-constant hazard-rate model for the duration of unemployment in single-interview samples of the stock of unemployed," *Economics Letters*, 51(2), 177 – 183.

RIDDER, G. (1987): "The sensitivity of duration models to misspecified unobserved heterogeneity and duration dependence," Working paper, Department of Econometrics, University of Groningen, Groningen.

ULYSSEA, G., AND D. SZERMAN (2006): "Job Duration and the Informal Sector in Brazil," Technical report, Instituto de Pesquisa Economica Aplicada.

WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Mass.

Table 4.5: Estimation Results Ignoring Stock Sampling (Sample CONTA)

	β_0	β_1	α	δ
True PS 1	-1.03	1	0.5	100
	-4.9745	2.7852	1.2328	0.9169
	(0.1268)	(0.1668)	(0.0437)	(0.1117)
True PS 2	-1.03	1	0.5	1
	-4.0924	2.6463	1.2876	0.1442
	(0.1196)	(0.2819)	(0.0558)	(0.0092)
True PS 3	-3.85	1	1	100
	-7.8336	2.0223	1.7462	1.5258
	(0.1983)	(0.1276)	(0.0531)	(0.2234)
True PS 4	-3.85	1	1	1
	-8.2289	2.2136	2.0740	0.1645
	(0.2930)	(0.2595)	(0.0942)	(0.0105)
True PS 5	-6.4	1	1.5	100
	-10.0088	1.6351	2.1664	3.0210
	(0.2708)	(0.1228)	(0.0678)	(0.7348)
True PS 6	-6.4	1	1.5	1
	-11.1958	1.8945	2.6151	0.2090
	(0.3909)	(0.2438)	(0.1078)	(0.0156)

NOTE: "True PS" refers to the true parameter set, see Table 4.2. Average point estimates and their standard deviation, in parenthesis, from the simulation samples. Estimation using the continuous-time data set.

Table 4.6: Estimation Results Accounting for Stock Sampling (Sample CONTA)

	β_0	β_1	α	δ
True PS 1	-1.03	1.00	0.5	100
	-1.0661	1.0394	0.5174	3.40E+06
	(0.1273)	(0.1267)	(0.0329)	(8.83E+06)
True PS 2	-1.03	1.00	0.5	1
	-0.0319	1.2571	0.5838	2.2542
	(0.1853)	(0.3404)	(0.0712)	(0.4550)
True PS 3	-3.85	1.00	1.0	100
	-3.9300	1.0348	1.0247	2.29E+06
	(0.1929)	(0.1112)	(0.0453)	(6.25E+06)
True PS 4	-3.85	1.00	1.0	1
	-3.2741	1.0332	1.0398	1.8546
	(0.2755)	(0.2108)	(0.0794)	(0.2878)
True PS 5	-6.4	1.00	1.5	100
	-6.4053	0.9992	1.5064	3.20E+06
	(0.2425)	(0.1041)	(0.0559)	(6.77E+06)
True PS 6	-6.4	1.00	1.5	1
	-5.9423	0.9988	1.5170	1.6518
	(0.3631)	(0.1831)	(0.0935)	(0.2366)

NOTE: "True PS" refers to the true parameter set, see Table 4.2. Average point estimates and their standard deviation, in parenthesis, from the simulation samples. Estimation using the continuous-time data set.

Table 4.7: Estimation Results with Imputed Elapsed Duration (Sample CONTB)

Imputation	β_0	β_1	α	δ	β_0	β_1	α	δ	β_0	β_1	α	δ
	Parameter Set 1				Parameter Set 3				Parameter Set 5			
	-1.03	1	0.5	100	-3.85	1	1	100	-6.4	1	1.5	100
E^L	-1.03	1.08	0.51	1.8E+06	-3.93	1.04	1.03	9.9E+25	-6.58	0.99	1.55	2.4E+06
	(0.14)	(0.14)	(0.04)	(4.8E+06)	(0.20)	(0.11)	(0.05)	(9.9E+26)	(0.28)	(0.11)	(0.06)	(4.5E+06)
$\frac{E^L+E^R}{2}$	-1.07	1.03	0.52	3.7E+06	-3.93	1.04	1.02	5.3E+29	-6.35	0.99	1.49	3.0E+06
	(0.12)	(0.12)	(0.03)	(7.9E+06)	(0.19)	(0.11)	(0.04)	(5.3E+30)	(0.25)	(0.11)	(0.06)	(5.7E+06)
E^R	-1.12	1.00	0.52	6.6E+06	-3.93	1.03	1.02	3.1E+13	-6.18	1.00	1.45	4.2E+26
	(0.11)	(0.11)	(0.02)	(9.0E+06)	(0.18)	(0.11)	(0.04)	(3.1E+14)	(0.26)	(0.11)	(0.06)	(4.2E+27)
	Parameter Set 2				Parameter Set 4				Parameter Set 6			
	-1.03	1	0.5	1	-3.85	1	1	1	-6.4	1	1.5	1
E^L	-0.03	1.42	0.63	1.86	-3.30	1.11	1.07	1.59	-5.97	1.03	1.53	1.46
	(0.19)	(0.38)	(0.08)	(0.35)	(0.31)	(0.23)	(0.09)	(0.28)	(0.40)	(0.19)	(0.11)	(0.24)
$\frac{E^L+E^R}{2}$	-0.03	1.23	0.57	2.35	-3.27	1.02	1.03	1.92	-5.94	1.00	1.52	1.69
	(0.18)	(0.33)	(0.07)	(0.48)	(0.27)	(0.21)	(0.08)	(0.29)	(0.36)	(0.18)	(0.09)	(0.24)
E^R	-0.01	1.07	0.52	2.97	-3.19	0.94	1.00	2.35	-5.87	0.96	1.49	1.98
	(0.18)	(0.30)	(0.06)	(0.66)	(0.24)	(0.19)	(0.06)	(0.32)	(0.33)	(0.17)	(0.08)	(0.25)

NOTE: Average point estimates and their standard deviation, in parenthesis, from the simulation samples. Estimation using the continuous-time data set.

Table 4.8: Estimation Results with Interval-Censored Residual Duration (Sample INTCA)

	β_0	β_1	α	δ
True PS 1	-1.03	1.00	0.5	100
	-1.0936	1.0517	0.5246	5.66E+06
	(0.1830)	(0.1387)	(0.0468)	(2.36E+07)
True PS 2	-1.03	1.00	0.5	1
	-0.2180	1.4543	0.6624	1.9859
	(0.4183)	(0.4418)	(0.1484)	(0.5450)
True PS 3	-3.85	1.00	1.0	100
	-3.9526	1.0388	1.0299	1.79E+06
	(0.2029)	(0.1134)	(0.0489)	(9.89E+06)
True PS 4	-3.85	1.00	1.0	1
	-3.3158	1.0390	1.0518	1.8259
	(0.3216)	(0.2108)	(0.0905)	(0.3207)
True PS 5	-6.4	1.00	1.5	100
	-6.4050	0.9996	1.5063	1.90E+06
	(0.2537)	(0.1058)	(0.0584)	(7.42E+06)
True PS 6	-6.4	1.00	1.5	1
	-5.9403	1.0000	1.5163	1.6548
	(0.3808)	(0.1836)	(0.0977)	(0.2385)

NOTE: "True PS" refers to the true parameter set, see Table 4.2. Average point estimates and their standard deviation, in parenthesis, from the simulation samples. Estimation using the continuous-time data set.

Table 4.9: Estimation Results with Interval-Censored Residual Duration and Imputed Elapsed Duration (Sample INTCB)

Imputation	β_0	β_1	α	δ	β_0	β_1	α	δ	β_0	β_1	α	δ
	Parameter Set 1				Parameter Set 3				Parameter Set 5			
	-1.03	1.00	0.5	100	-3.85	1.00	1.0	100	-6.4	1.00	1.5	100
E^L	-1.03	1.09	0.51	1.1E+06	-3.96	1.04	1.03	8.3E+09	-6.54	0.98	1.54	7.4E+08
	(0.21)	(0.15)	(0.06)	(3.9E+06)	(0.21)	(0.12)	(0.05)	(8.3E+10)	(0.35)	(0.11)	(0.08)	(7.1E+09)
$\frac{E^L+E^R}{2}$	-1.11	1.05	0.53	1.9E+10	-3.95	1.04	1.03	5.6E+05	-6.35	0.99	1.49	7.5E+05
	(0.17)	(0.13)	(0.04)	(1.9E+11)	(0.20)	(0.11)	(0.05)	(1.7E+06)	(0.26)	(0.11)	(0.06)	(1.8E+06)
E^R	-1.17	1.02	0.54	5.9E+06	-3.95	1.04	1.03	1.2E+06	-6.16	1.00	1.45	7.3E+06
	(0.18)	(0.12)	(0.04)	(2.3E+07)	(0.19)	(0.11)	(0.04)	(3.7E+06)	(0.26)	(0.11)	(0.06)	(6.4E+07)
	Parameter Set 2				Parameter Set 4				Parameter Set 6			
	-1.03	1.00	0.5	1	-3.85	1.00	1.0	1	-6.4	1.00	1.5	1
E^L	-0.25	1.78	0.75	1.59	-3.35	1.13	1.08	1.56	-5.93	1.02	1.52	1.49
	(0.51)	(0.54)	(0.20)	(0.48)	(0.38)	(0.23)	(0.11)	(0.32)	(0.46)	(0.20)	(0.12)	(0.31)
$\frac{E^L+E^R}{2}$	-0.21	1.40	0.65	2.08	-3.31	1.02	1.05	1.89	-5.94	1.00	1.52	1.69
	(0.40)	(0.42)	(0.14)	(0.57)	(0.31)	(0.21)	(0.09)	(0.32)	(0.38)	(0.18)	(0.10)	(0.24)
E^R	-0.12	1.13	0.56	2.78	-3.22	0.94	1.00	2.33	-5.87	0.96	1.49	1.98
	(0.34)	(0.34)	(0.11)	(0.79)	(0.27)	(0.19)	(0.07)	(0.34)	(0.35)	(0.17)	(0.09)	(0.26)

NOTE: Average point estimates and their standard deviation, in parenthesis, from the simulation samples. Estimation using monthly, interval-censored residual duration and interval-censored elapsed duration for spells that started more than 52 weeks before the stock sampling date.

Table 4.10: Estimation Results Monthly and Interval-Censored Duration Data: Imputed Elapsed Duration (Sample INTCC)

Imputation	β_0	β_1	α	δ	β_0	β_1	α	δ	β_0	β_1	α	δ
	Parameter Set 1				Parameter Set 3				Parameter Set 5			
	-1.03	1	0.5	100	-3.85	1	1	100	-6.4	1	1.5	100
E^L	-1.71	1.22	0.66	4.09E+05	-3.93	1.04	1.03	2.74E+08	-6.34	0.98	1.50	1.47E+10
	(0.14)	(0.15)	(0.05)	(3.1E+06)	(0.18)	(0.11)	(0.04)	(2.7E+09)	(0.26)	(0.10)	(0.06)	(1.1E+11)
$\frac{E^L+E^R}{2}$	-1.04	1.07	0.52	7.83E+10	-3.96	1.04	1.03	3.48E+06	-6.35	0.99	1.49	1.68E+08
	(0.36)	(0.31)	(0.11)	(7.8E+11)	(0.20)	(0.11)	(0.05)	(2.5E+07)	(0.26)	(0.11)	(0.06)	(1.7E+09)
E^R	-0.76	1.00	0.47	5.63E+06	-3.97	1.04	1.03	1.12E+06	-6.33	1.00	1.48	2.98E+05
	(0.14)	(0.11)	(0.02)	(1.3E+07)	(0.22)	(0.11)	(0.05)	(3.7E+06)	(0.28)	(0.11)	(0.06)	(8.4E+05)
	Parameter Set 2				Parameter Set 4				Parameter Set 6			
	-1.03	1	0.5	1	-3.85	1	1	1	-6.4	1	1.5	1
E^L	-1.23	1.25	0.90	1.23	-3.33	1.05	1.06	1.61	-5.65	0.99	1.46	1.58
	(0.25)	(0.35)	(0.13)	(0.23)	(0.27)	(0.21)	(0.08)	(0.27)	(0.37)	(0.19)	(0.10)	(0.25)
$\frac{E^L+E^R}{2}$	0.26	1.73	0.58	1,273.661†	-3.31	1.03	1.05	1.89	-5.95	1.00	1.52	1.69
	(0.66)	(0.63)	(0.21)	(12,711.35)	(0.32)	(0.21)	(0.09)	(0.33)	(0.39)	(0.19)	(0.10)	(0.25)
E^R	3.50	1.52	0.43	277,464.3‡	-3.20	0.99	1.01	2.28	-6.12	0.99	1.54	1.89
	(8.63)	(5.15)	(0.32)	(1,393,749)	(0.33)	(0.20)	(0.09)	(0.38)	(0.47)	(0.19)	(0.11)	(0.35)

NOTE: Average point estimates and their standard deviation, in parenthesis, from the simulation samples. Estimation using monthly, interval-censored residual duration and interval-censored elapsed duration for spells that started more than 52 weeks before the stock sampling date.

† The large value in this average is because for one sample the estimated δ was very large. Taking the average over the remaining 99 samples gives an average of 2.53 (with a standard deviation of 1.25).
‡ The large value in this average is because for 25 samples the estimated δ was very large. Taking the average over the remaining 75 samples gives an average of 3.57 (with a standard deviation of 2.37).

Chapter 5

Conclusion

Even though the idea of accumulation of human capital in the informal sector has been suggested in previous studies, no study has attempted to verify this possibility. Chapter 2 provides evidence and arguments that suggest that informal sector jobs might indeed provide training opportunities for young less-educated workers. Chapter 3 goes further in trying to understand the role informal jobs play in the careers of less-educated workers. In addition to the role of informal jobs as providers of training opportunities, this chapter also investigates a second role that is rarely mentioned in the literature: the role of informal jobs as a screening mechanism that allows employers to learn the skills of young less-educated workers.

The research strategy followed in Chapter 3 is to build a model of the informal and formal sectors that can separately accommodate two roles of informal

jobs: human capital accumulation and screening. Each of these models produces different implications on the shape of the hazard function from the informal to the formal sector. These differences in the hazard function are used to determine whether informal sector jobs play the role of skill formation or screening in the early careers of less-educated workers. The estimated hazard function is consistent with the implications of the screening model, which indicates that the informal sector has an important role by screening young less-educated workers entering the labor market. It is stressed that his result does not rule out the provision of human capital in the informal sector.

Finally, the empirical analysis in Chapter 3 required exploration of the finite-sample properties of the estimators of the hazard functions when the data do not have the properties required by estimation methods suggested in the literature. It is argued that this departure from the typical stock sampling data occurs often in practice, and for this reason, this problem cannot be ignored. The problem in the data is that the exact starting times of spells are not always observed. Instead, only an interval containing the exact starting time is observed. Chapter 4 provides a Monte Carlo analysis to explore the finite-sample properties of estimators using different methods to impute the interval-censored starting time. The results from the Monte Carlo analysis indicated that using the mid-point of the interval yields satisfactory results when working with duration data obtained from surveys that are implemented as rotating panels, such as the one used in the empirical analysis of Chapter 3,

hence these were used in the empirical analysis of that chapter.

Appendix A

Appendix for Chapter 2

A.1 Wage Imputations

The information on earnings and working hours collected by the ENEU refers to the job held during the week previous to the interview, which is called the *reference week*. However, if the respondent did not attend work during the reference week, this information is missing. This section explains the methodology and criteria used in this paper to impute the respondent's earnings and working hours when they are missing.

If the respondent was absent from work during the reference week, but stated to have a job, then the ENEU proceeds to determine why the respondent did not attend work. Some of the reasons why the respondent might have been absent from work are: vacation, sickness and recovery, strike, lack of production inputs, and work season ended. In such cases, the ENEU collects information on *usual earnings* and *usual working hours*. The information on usual earnings and usual working hours is used in this paper to impute the missing earnings and working hours only when the respondent declared to be absent from work due to vacation or due to sickness and recovery. Then, this information is used to compute a measure of usual hourly earnings, which in turn is used to impute the missing hourly earnings.

Finally, only those measures of usual earnings that satisfy certain criteria are used to impute the missing hourly earnings. The criteria is to compare the

measure of usual hourly earnings against hourly earnings from the previous and subsequent interviews, and to impute missing hourly earnings whenever the measure of usual hourly earnings are within one standard deviation from the previous, or the subsequent, measure of hourly earnings. The standard deviation is obtained with respect to the hourly earnings observations of each respondent. About 1% of the measures of hourly earnings in the final sample are the imputed hourly earnings.

As mentioned in the description of the sample, the top and bottom 1% of the hourly earnings are dropped from the sample. Only after the top and bottom 1% are dropped are missing hourly earnings imputed.

Appendix B

Appendix for Chapter 3

B.1 Wages in the Model

The surplus sharing rule implies that:

$$w_F(x,p) \quad \text{is such that:} \qquad W_F(x,p) - U(p) = \frac{\beta}{1-\beta}\big[J_F(x,p) - V_F\big]$$
$$w_I(p) \quad \text{is such that:} \qquad W_I(p) - U(p) = \frac{\beta}{1-\beta}\big[J_I(p) - V_I\big]$$

where in equilibrium, free entry implies that $V_F = 0$ and $V_I = 0$.

Wages in the Baseline Model:

$$w_F(x,p) = \beta\big(px - \delta D\big) + (1-\beta)\tilde{r}U(p)$$
$$w_I(p) = \beta p_I + (1-\beta)\left(\tilde{r}U(p) - \beta m(\theta_F)\int_{Q(p)}^{1} S_F(s,p)dG(s)\right)$$

Wages in the Human Capital Model:

$$w_F(x,p) = \beta\big(px - \delta D\big) + (1-\beta)\Big(\tilde{r}U(p) - \kappa\big[U(p_H) - U(p)\big]\Big)$$
$$w_I(p) = \beta p_I + (1-\beta)\left(\tilde{r}U(p) - \kappa\big[U(p_H) - U(p)\big] - \beta m(\theta_F)\int_{Q(p)}^{1} S_F(s,p)dG(s)\right)$$

Wages in the Learning Model:

$$w_F(x) = \beta\Big(\bar{p}x - \delta D - \sigma\phi\Gamma_L(x)D\Big) + (1-\beta)\Big(\tilde{r}U - \sigma[\phi U(p_L) + (1-\phi)U(p_H) - U]\Big)$$

$$w_I = \beta p_I + (1-\beta)\left(\tilde{r}U - \sigma[\phi U(p_L) + (1-\phi)U(p_H) - U] - \beta m(\theta_F)\int_Q^1 S_F(x)dG(x)\right)$$

B.2 Proofs

B.2.1 Proof of Lemma 1

Note that $S_I(p) = \dfrac{J_I(p)}{1-\beta}$. In the proof we replace $S_I(p)$ with $J_I(p)/(1-\beta)$ in (3.9). Consider the following result which proves to be useful in the proof of Lemma 1.

Lemma 2. *An upper bound for* $\big[Q(p) - C(p)\big]$ *is* $\dfrac{\tilde{r}+\delta}{\tilde{r}+\delta+\mu(p)+m(\theta_I)\beta}\left(\dfrac{p_I - z}{p}\right)$.

Proof. First, note that:

$$\begin{aligned}
J_I(p) &= \frac{1-\beta}{\tilde{r}+\delta+\mu(p)}\left(p_I - \tilde{r}U(p) + m(\theta_F)\int_{Q(p)}^1 [W_F(x,p) - U(p)]dG(x)\right) \\
&< \frac{1-\beta}{\tilde{r}+\delta+\mu(p)}\left(p_I - \tilde{r}U(p) + m(\theta_F)\int_{C(p)}^1 [W_F(x,p) - U(p)]dG(x)\right) \\
&= \frac{1-\beta}{\tilde{r}+\delta+\mu(p)}\Big(p_I - z - m(\theta_I)[W_I(p) - U(p)]\Big) \\
&= \frac{1-\beta}{\tilde{r}+\delta+\mu(p)}\Big(p_I - z - m(\theta_I)\frac{\beta}{1-\beta}J_I(p)\Big)
\end{aligned}$$

And so, $J_I(p) < \dfrac{1-\beta}{\tilde{r}+\delta+\mu(p)+\beta m(\theta_I)}(p_I - z)$. Since $Q(p) - C(p) = \left(\dfrac{\tilde{r}+\delta}{p}\right)\dfrac{J_I(p)}{1-\beta}$, the result follows. □

Next, we proceed to prove Lemma 1.

Proof. Note that once we substitute equilibrium wage equations and use the surplus sharing rules to substitute for unknown value functions, equations (3.1), (3.8), and (3.9) represent a system of three equations with three unknowns and one parameter:

$$F(\tilde{r}U, C, Q; p) = 0. \tag{B.1}$$

Note that we treat $\Omega = (\delta, p_I, m(\theta_I), m(\theta_F), r, z, D, \beta)$ as given because Ω does not change when the parameter p changes. Linearizing (B.1), we get:

$$\begin{aligned}
(1-A)d(\tilde{r}U) + (N+AK)dC + (A(L-E))dQ - (AH+M)dp &= 0 \\
-\frac{1}{p}d(\tilde{r}U) + dC + \frac{C}{p}dp &= 0 \\
Bd(\tilde{r}U) + (BK-1)dC + (1+B(L-E))dQ - \left(BH - \frac{Q-C}{p}\right)dp &= 0
\end{aligned}$$

where:

$$\begin{aligned}
A &= \frac{m_I\beta}{\tilde{r}+\delta+\mu(p)}, & K &= \frac{\mu(p)\beta}{\tilde{r}+\delta}p, \\
B &= \frac{\tilde{r}+\delta}{p(\tilde{r}+\delta+\mu(p))}, & L &= \frac{m_F\beta}{\tilde{r}+\delta}p(Q-C)g(Q), \\
E &= m_F g(Q)\frac{J_I}{1-\beta}, & M &= \frac{m_F\beta}{\tilde{r}+\delta}\int_C^1 (x-C)dG(x), \\
H &= \frac{m_F\beta}{\tilde{r}+\delta}\int_Q^1 (x-C)dG(x), & N &= \frac{m_F\beta}{\tilde{r}+\delta}p[1-G(C)]
\end{aligned}$$

Note that for ease of exposition we denoted $m(\theta_j) = m_j$ for $j \in \{F, I\}$. By the Implicit Function Theorem and using Cramer's rule, we can derive dC/dp, which is given by:

$$\frac{dC}{dp} = \frac{A(L-E)(Q-C) + pB(L-E)(M-C) + pA(H-C) + p(M-C)}{p\big[(L-E)(BN+A+pB) + AK + Ap + N + p\big]}.$$

It is straightforward to show that all the terms in the numerator, except for the second one, are negative. Ignore the third term, which we know is negative, then adding the first, second, and fourth terms in the numerator, and after some algebra, we get:

$$p\left(\frac{\delta D+z}{p}\right)\left[\frac{m_F(1-\beta)}{\tilde{r}+\delta+\mu(p)}(Q-C)g(Q)-1\right]-\frac{m_I\beta}{\tilde{r}+\delta}p(Q-C)$$

which is negative if the term in square brackets is negative. Using Lemma 2 to bound the term in square brackets from above we get:

$$\left[\frac{m_F(1-\beta)}{\tilde{r}+\delta+\mu(p)}(Q-C)g(Q)-1\right]<\left(\frac{\tilde{r}+\delta}{\tilde{r}+\delta+\mu(p)}\right)\left(\frac{(1-\beta)m_Fg(Q)}{\tilde{r}+\delta+\mu(p)+m_I\beta}\right)\left(\frac{p_I-z}{p}\right)-1.$$

Notice that because $\mu(p)>0$, we can further bound the term on the right hand side of the inequality from above. Then, a sufficient condition for the numerator to be negative is that:

$$\left(\frac{(1-\beta)m_Fg(Q)}{\tilde{r}+\delta+\mu(p)+m_I\beta}\right)\left(\frac{p_I-z}{p}\right)<1.$$

Let $\eta=\dfrac{1}{1-\beta}\left(\dfrac{p_L}{p_I-z}\right)$. Rearranging this condition we get:

$$m_F\Big(g(Q)-\eta[1-G(Q)]\Big)<\eta(\tilde{r}+\delta)+\eta m_I\beta.$$

Finally, since $m_I>0$, $m_F<1$, and $Q\in[0,1]$, a stronger condition that does not depend on endogenous variables is:

$$\forall x\in[0,1]\quad\Big(g(x)-\eta\int_x^1 g(u)du\Big)<\eta(\tilde{r}+\delta)\qquad\text{(CDN 1)}$$

Since the third term is negative, (CDN 1) is a sufficient condition for the numerator to be negative. Now, we focus on the denominator of dC/dp. It is

straightforward to show that $(L-E)(BN+A+pB)<0$, and using Lemma 2 again, we can bound from above the absolute value of the this term:

$$\Big|(L-E)(BN+A+pB)\Big| < g(Q)\frac{m_F(1-\beta)(p_I-z)}{\tilde{r}+\delta+\mu(p)+m_I\beta}\left[\frac{m_I\beta+m_F\beta[1-G(Q)]+\tilde{r}+\delta}{\tilde{r}+\delta+\mu(p)}\right]. \tag{B.2}$$

The other term in the denominator is positive and it is given by:

$$(AK+Ap+N+p) = p\left[\left(\frac{m_I\beta}{\tilde{r}+\delta+\mu(p)}\right)\left(\frac{\mu(p)\beta+\tilde{r}+\delta}{\tilde{r}+\delta}\right)+\frac{m_F\beta[1-G(C)]}{\tilde{r}+\delta}+1\right]. \tag{B.3}$$

Next, we compare (B.2) and (B.3). Using the sufficient condition (CDN 1) we can show that $p > g(Q)\frac{m_F(1-\beta)(p_I-z)}{\tilde{r}+\delta+\mu(p)+m_I\beta}$, so the outer term is higher for (B.3). Finally, it is straightforward to show that the term in square brackets is also higher in (B.3) than the term in square brackets in (B.2), so that the denominator is positive. As a result, $dC/dp<0$.

Now, we apply the Implicit Function Theorem and use Cramer's rule again to derive dQ/dp, which is given by:

$$\frac{dQ}{dp} = \frac{pB\big[H(N+p)-M(K+p)+C(K-N)\big]+p\big[(M-Q)+A(H-Q)\big]+(N+AK)(C-Q)}{p\Big[(L-E)(BN+A+pB)+AK+Ap+N+p\Big]}.$$

We already proved that under certain parameter conditions the denominator of dQ/dp is positive. Then it just remain to show that the numerator is negative. It is straightforward to show that the second and third terms of the numerator are negative. To show that the first term is negative, note that $M>H$ so:

$$pB[H(N+p)-M(K+p)+C(K-N)] < pB[M(N+p)-M(K+p)+C(K-N)]$$

$$
\begin{aligned}
&= pB[MN - MK + C(K - N)] \\
&= pB(N - K)[M - C] \\
&< 0
\end{aligned}
$$

where the last inequality from the fact that $C > M$. As a result, $dQ/dp < 0$. And this completes the proof.

□

B.2.2 Proofs of the Shape of the Unconditional Hazard Rates

Before proving Propositions 2, 4, and 6, consider the following result about the unconditional hazard rate. The proof of Lemma 3 follows the arguments of **?**, chap. 4.

Lemma 3. *Let $\lambda(t|p)$ be the hazard rate conditional on worker skill level, and $\lambda'(t|p) = \partial\lambda(t|p)/\partial t$. Let ϕ_I be the probability that $p = p_L$ in the informal sector. Then, the unconditional hazard rate and its derivative are given by:*

$$\lambda(t) = \gamma(t)\lambda(t|p_L) + [1 - \gamma(t)]\lambda(t|p_H)$$

$$\lambda'(t) = \gamma'(t)\big[\lambda(t|p_L) - \lambda(t|p_H)\big] + \gamma(t)\lambda'(t|p_L) + [1 - \gamma(t)]\lambda'(t|p_H)$$

where $\gamma(t) = \frac{1}{1+\eta(t)}$, $\eta(t) = \left(\frac{1-\phi_I}{\phi_I}\right) e^{-[\Lambda(t|p_H) - \Lambda(t|p_L)]}$, $\Lambda(t|p) = \int_0^t \lambda(s|p)ds$, *and* $\eta'(t) = \eta(t)\big[\lambda(t|p_L) - \lambda(t|p_H)\big]$.

Proof. The conditional survivor function is given by $S(t|p) = e^{-\Lambda(t|p)}$. Then, the unconditional survivor function is given by $S(t) = \phi_I e^{-\Lambda(t|p_L)} + (1 - \phi_I)e^{-\Lambda(t|p_H)}$,

and the unconditional hazard is given by $\lambda(t) = -d \ln S(t)/dt$, then by the First Fundamental Theorem of Calculus:

$$\lambda(t) = \frac{\phi_I \lambda(t|p_L) e^{-\Lambda(t|p_L)} + (1-\phi_I)\lambda(t|p_H)e^{-\Lambda(t|p_H)}}{\phi_I e^{-\Lambda(t|p_L)} + (1-\phi_I)e^{-\Lambda(t|p_H)}} = \gamma(t)\lambda(t|p_L) + [1-\gamma(t)]\lambda(t|p_H)$$

and

$$\gamma(t) = \frac{\phi_I e^{-\Lambda(t|p_L)}}{\phi_I e^{-\Lambda(t|p_L)} + (1-\phi_I)e^{-\Lambda(t|p_H)}} = \frac{1}{1+\eta(t)}$$
$$\eta(t) = \left(\frac{1-\phi_I}{\phi_I}\right) e^{-[\Lambda(t|p_H) - \Lambda(t|p_L)]},$$

so that $\eta(t) > 0$. $\lambda'(t)$ is straightforward and applying the First Fundamental Theorem of Calculus again we have:

$$\eta'(t) = \eta(t)[\lambda(t|p_L) - \lambda(t|p_H)].$$

□

Proof of Proposition 2

Proof. From Lemma 3 and Proposition 1 we have that:

$$\eta'(t) = \eta(t)[\mu(p_L) - \mu(p_H)] < 0,$$
$$\gamma'(t) = -\gamma(t)^2\eta(t)[\mu(p_L) - \mu(p_H)] > 0, \text{ and}$$
$$\lambda'(t) = \gamma'(t)[\mu(p_L) - \mu(p_H)] < 0.$$

□

Proof of Proposition 4

Proof. From Lemma 3 and Proposition 3 we have that:

$$\eta'(t) = \eta(t)(1-\kappa)^t\big[\mu(p_L) - \mu(p_H)\big] < 0,$$

$$\gamma'(t) = -\gamma(t)^2\eta(t)(1-\kappa)^t\big[\mu(p_L)-\mu(p_H)\big] > 0, \text{ and}$$
$$\lambda'(t) = \gamma(t)(1-\kappa)^t[\mu(p_L)-\mu(p_H)]^2\left[\frac{\ln(1-\kappa)}{\mu(p_L)-\mu(p_H)} - \gamma(t)\eta(t)(1-\kappa)^t\right],$$

where each term in the square brackets is positive. However, the first term in the square brackets is constant while the second one decreases with time. To see this, define $\Phi(t) = \gamma(t)\eta(t)(1-\kappa)^t$, then it is easy to check that

$$\Phi'(t) = \frac{\eta'(t)}{[1+\eta(t)]^2}(1-\kappa)^t + \frac{\eta(t)}{1+\eta(t)}(1-\kappa)^t\ln(1-\kappa) < 0$$

where negativity follows from $\eta'(t) < 0$ and $\kappa \in (0,1)$. Evaluating $\Phi(t)$ at $t=0$ we find that $\Phi(0) = 1-\phi_I$, therefore:

(i) if $\ln(1-\kappa)/[\mu(p_L)-\mu(p_H)] > (1-\phi_I)$, then the term in square brackets is always positive, and

(ii) if $\ln(1-\kappa)/[\mu(p_L)-\mu(p_H)] < (1-\phi_I)$, then the term in square brackets is initially negative, but becomes eventually positive, so that $\lambda(t)$ decreases initially, but eventually increases.

□

Proof of Proposition 6

Proof. From Lemma 3 and Proposition 5 we have that

$$\eta'(t) = \eta(t)[1-(1-\sigma)^t]\big[\mu(p_L)-\mu(p_H)\big] < 0$$
$$\gamma'(t) = -\gamma(t)^2\eta(t)[1-(1-\sigma)^t]\big[\mu(p_L)-\mu(p_H)\big] > 0, \text{ and}$$
$$\lambda'(t) = -\gamma(t)^2\eta(t)[1-(1-\sigma)^t]^2\big[\mu(p_L)-\mu(p_H)\big]^2$$

$$+ (1-\sigma)^t \ln(1-\sigma)\big[\bar{\mu} - \phi\mu(p_L) - (1-\phi)\mu(p_H)\big].$$

Note that in the definition of $\eta(t)$ in Lemma 3, ϕ replaces ϕ_I. Inspection of $\lambda'(t)$ reveals that for low values of t, the second term dominates but it is eventually overtaken by the first term, much more faster the higher σ is. Next, evaluating $\lambda'(t)$ at $t = 0$, we find

$$\lambda'(0) = \ln(1-\sigma)\big[\bar{\mu} - \phi\mu(p_L) - (1-\phi)\mu(p_H)\big].$$

Therefore, if the term in square brackets is:

(i) Positive, then the hazard is monotonically decreasing.

(ii) Negative, then the hazard increases initially, but eventually decreases.

□

B.3 Minimization Algorithm to Find Parameters of the Employer Learning Model

The estimated hazard suggest starting values for $\big(\bar{\mu}, \mu(p_L), \mu(p_H)\big)$. In particular, by Condition 3, $Q(p_H) < Q < Q(p_L)$, and so $\mu(p_L) < \bar{\mu} < \mu(p_H)$. This is because at $t = 0$ the hazard must equal $\bar{\mu}$ and for longer durations the hazard must equal $\mu(p_L)$. However, the estimated hazard in Figure 3.5 suggests that $Q \approx Q(p_L)$. Then, we set $\mu(p_L) = 0.99 \cdot \bar{\mu}$, so that $\mu(p_L)$ is arbitrarily close to, but below $\bar{\mu}$, and use $\exp(\bar{x}'\hat{\beta})\hat{\lambda}_1 = 0.03$ as a starting value for $\bar{\mu}$. Similarly, we know that $\mu(p_H)$ must be higher than the maximum of the hazard function, then we use $\exp(\bar{x}'\hat{\beta})\hat{\lambda}_2 = 0.3$ as a starting value for $\mu(p_H)$. The estimated hazard does not provide much information to select starting values for (σ, ϕ). Hence we use different starting values given by $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ for each parameter. This gives a total of 25 different starting values, in all cases we use $\overline{T} = 50$. For all starting values, the resulting vector of parameters is: $\bar{\mu} = 0.05$, $\mu(p_L) = 0.0495$, $\mu(p_H) = 1.0$, $\phi = 0.4833$, and $\sigma = 0.1478$.

Currirculum Vitae

Name:	Javier Cano Urbina
Place of Birth:	Tampico, Mexico
Year of Birth:	1975
Post-Secondary Education and Degrees:	Universidad Autónoma de Nuevo León Monterrey, NL, Mexico 1995–1999 B.A. Economics, *cum laude*
	University of Rochester Rochester, NY, USA 2003–2005 M.A. Economics
	The University of Western Ontario London, Ontario 2006–2012 Ph.D. Economics
Honors and Awards:	Summer Research Fellow Banco de México 2010
	Graduate Fellowship & Tuition Scholarship The University of Western Ontario 2009-2010
	Graduate Fellowship & Tuition Scholarship CONACYT 2004-2009
Related Work Experience:	*Graduate Teaching Assistant* The University of Western Ontario 2008–2010
	Instructor The University of Western Ontario 2011
	Research Assistant The University of Western Ontario 2007–2011