Puggioni, Daniela et al.

Working Paper

Inequality, income dynamics, and transitions of Mexican workers

Working Papers, No. 2022-14

Provided in Cooperation with:
Bank of Mexico, Mexico City

Suggested Citation: Puggioni, Daniela et al. (2022) : Inequality, income dynamics, and transitions of Mexican workers, Working Papers, No. 2022-14, Banco de México, Ciudad de México

This Version is available at:
https://hdl.handle.net/10419/273670

Banco de México

Working Papers

N° 2022-14

Inequality, Income Dynamics, and Transitions of Mexican Workers

Daniela Puggioni
Banco de México

Mariana Calderón
Banco de México

Alfonso Cebreros
Banco de México

León Fernández
Banco de México

José A. Inguanzo
UCLA

David Jaume
Banco de México

November 2022

Documento de Investigación
2022-14

Working Paper
2022-14

Inequality, Income Dynamics, and Transitions of Mexican Workers

Daniela Puggioni[†]
Banco de México

Mariana Calderón[‡]
Banco de México

Alfonso Cebreros[§]
Banco de México

León Fernández[**]
Banco de México

José A. Inguanzo[±]
UCLA

David Jaume[⊤]
Banco de México

Abstract: We characterize the salient features of the distribution of earnings and earnings changes of formal workers in Mexico using social security records for the period 2005-2019. We find strong evidence of deviations from normality of these distributions. Comparing the results obtained with administrative data and household survey data suggests that this latter source of information is inadequate to fully capture the evolution of inequality and the properties of earnings changes. We also study the impact of transitions out of and back into formal employment on wages earned in the formal sector and the effect of early exposure to informality on future earnings. We document that workers who exit formal employment experience a significant wage penalty upon re-entry and that having the first job in the informal sector has a negative and significant impact on future earnings.

Keywords: Earnings dynamics, higher-order earnings risk, inequality, worker transitions, informal labor markets.

JEL Classification: E24, J24, J31, J46.

Resumen: Analizamos las características principales de la distribución de los ingresos y de los choques a los ingresos de los trabajadores formales en México utilizando registros de seguridad social para el periodo 2005-2019 y encontramos evidencia de desviaciones de la normalidad. Una comparación entre los resultados obtenidos con datos administrativos y datos de encuestas de hogares sugiere que esta última fuente de información es inadecuada para capturar con precisión la evolución de la desigualdad y de los choques a los ingresos. También estudiamos el impacto de las transiciones desde y hacia el empleo formal sobre los salarios ganados en el sector formal y el efecto de la exposición temprana a la informalidad sobre los ingresos futuros. Documentamos que los trabajadores que salen del empleo formal experimentan una penalización salarial significativa al volver a ingresar y que tener el primer trabajo en el sector informal tiene un impacto negativo en los ingresos futuros.

Palabras Clave: Dinámicas del ingreso, riesgo de orden superior del ingreso, desigualdad, transiciones de los trabajadores, mercados laborales informales.

Dirección General de Investigación Económica. Emails: † dpuggionih@banxico.org.mx; ‡ mcalderon@banxico.org.mx; § carlos.cebreros@banxico.org.mx; ** lfernandezb@banxico.org.mx.

± University of California at Los Angeles. Email: joseinguanzo@ucla.edu.

⊤ Dirección General de Estabilidad Financiera. Email: djaumep@banxico.org.mx.

1 Introduction

There is a large and growing literature studying the salient features of the distribution and dynamics of earnings and life-time income that has focused primarily on advanced economies such as the United States, the United Kingdom, Germany, and Spain, to name a few (see, for example, Bonhomme and Hospido [2017] and Guvenen, Karahan, Ozkan, and Song [2021]). Comparatively, little attention has been devoted to emerging and developing countries. The goal of this paper is to contribute to this literature along two dimensions. First, we characterize the defining properties of the distribution of earnings and of transitory earnings changes and the extent to which they display deviations from normality using administrative records of workers employed in the formal sector. We explore how these properties vary across genders and age groups, and along the income distribution.[1] Second, we analyze the effect of transitions out of and back into formal employment on the earnings of workers and the role of early exposure to informality in shaping the dynamics of earnings in the long run. On both fronts, to the best of our knowledge, we are the first to explore these issues in the Mexican context.

Regarding the distribution of earnings, we find that overall inequality (P90–P10 dispersion) remained fairly stable during the 2005–2015 decade but started to fall after 2016. We also find that inequality increased during the period of the 2008–2009 financial crisis, more so for men than for women, and for younger workers than for older ones. Initial inequality (P90–P10 dispersion at age 25) has also been fairly stable between 2005 and 2019, with the exception of a marginal increase registered during the financial crisis. Within the limited time span covered by our data, we observe relatively stable earnings mobility patterns with lower income workers moving upward in the permanent income distribution and higher income workers being more likely to move downward.

With respect to the distribution of one-year earnings changes, we encounter evidence of

[1]This paper is an extended version of our paper *Inequality, Income Dynamics, and Worker Transitions: The Case of Mexico*, that will appear in a special issue of Quantitative Economics, along with other papers from all the countries that participate in the Global Income Dynamics Project. The statistics reported in this paper, as well as several additional ones that we computed on the administrative data but that are not explicitly included here, will be publicly available through the Global Repository of Income Dynamics (GRID), that will house harmonized and comparable statistics for all participating countries (https://www.grid-database.org). The administrative data (from the Instituto Mexicano del Seguro Social, IMSS) used in this paper are confidential and were made available through the Econlab at Banco de México. Inquiries regarding the terms and conditions for accessing these data should be directed to econlab@banxico.org.mx. We are grateful to the organizers of the Global Income Dynamics Project, Fatih Guvenen, Gianluca Violante, and Luigi Pistaferri. We also thank the participants of the Global Income Dynamics conferences (at Stanford and University of Minnesota), the 55th Annual Canadian Economics Association Meetings, and the Informal Seminar at Banco de México's research department for insightful comments and suggestions. We acknowledge Alejandro Trujillo for outstanding research assistance.

strong deviations from normality whose extent varies across income levels, age groups and genders. Similar to the findings of Guvenen et al. [2021], we document very high kurtosis across all demographic groups, more so for women, even though it has been steadily decreasing for both men and women since 2009. While the distribution of transitory earnings innovations is asymmetric, we find that skewness can be positive or negative depending on age, gender, and income. This contrasts with what Guvenen et al. [2021] document for the United States where this distribution is negatively skewed for all workers.

To document these facts regarding the distribution of earnings and earnings changes for Mexican workers, our analysis relies on social security records from the Instituto Mexicano del Seguro Social (IMSS) covering the period from 2005 to 2019 and the universe of (private sector) formal workers. Unlike previous work based on Mexican survey data (see, for example, Binelli and Attanasio [2010]), our social security records have large sample sizes and allow for following workers continuously throughout their employment history in the formal sector.

Social security data have the advantage of providing accurate, reliable, and consistent information on millions of Mexican workers,[2] but they also suffer from important limitations. First, as it is commonly the case with administrative data, earnings are bottom and top coded. Second, it is not possible to separate wage effects from labor-supply effects on earnings since no information on the number of hours worked or the full versus part-time status of a worker is available. Third, Mexican social security records do not provide important information on worker characteristics such as educational attainment or occupation of employment, precluding the possibility of exploring how the distribution of earnings or earnings changes varies along these dimensions. The main limitation of our data, however, is that they only cover workers employed in the formal private sector. Hence, while the power of millions of administrative records can be exploited to establish key features of the distribution of earnings and earnings changes for workers employed in the formal sector, this source of information has nothing to say about these distributions for informal workers, who constitute a large fraction (above 50%) of the Mexican labor market.

Motivated by this lack of information on the informal sector, we explore additional data sources. In particular, we contrast some of our baseline results obtained with the administrative data with results obtained selecting a comparable sample from household surveys. Even if several of the key insights from our analysis, especially those regarding inequality, are maintained across different data sources and for both formal and informal workers, we also

[2]It has been documented that employers used to under-report wages to IMSS, such practice, however, has substantially declined since the 1997 reform of the pension system in Mexico. Hence, under-reporting does not appear to be a problem for the period covered in our study. See Kumler, Verhoogen, and Frías [2020].

find important differences. In our context, these differences are likely associated with the fact that the share of workers not responding to income-related questions in the household survey has been growing over time and that this missing information is non-random: non-response is mostly concentrated among formal and highly educated workers who tend to belong to the upper part of the income distribution (as also documented in Campos Vázquez [2013] and Campos Vázquez and Lustig [2019]) . This comparison suggests that household surveys may be inadequate to fully capture the evolution of inequality and the properties of earnings changes.

Given that one of the consequences of a large informal sector is that many workers maintain a tenuous attachment to formal employment, we study how time away from formal employment shapes earnings dynamics. We provide empirical evidence that workers who temporarily exit formal employment experience a wage penalty upon re-entry, with earnings taking up to three years to gradually grow toward pre-exit levels. This suggests that the ability to maintain a continuous attachment to the formal sector is likely to be an important source of heterogeneity in the distribution of earnings changes across Mexican workers.

Lastly, we exploit a special module of the Mexican household survey that tracks individual labor market trajectories, and find that, when a worker's first job is in the informal sector, this is associated with a long-lasting negative effect on his/her earnings. This last set of results indicates that another important determinant of the heterogeneity in earnings levels and earnings changes that we document based on social security records can be traced back to initial conditions and the extent to which informality shaped workers' early labor market experiences.

Related literature. Methodologically, the first half of this paper is related to the work of Arellano, Blundell, and Bonhomme [2017], Guvenen, Kaplan, Song, and Weidner [2017], and Guvenen et al. [2021], among others. These authors propose non-parametric methods to characterize key features of the distribution of earnings shocks and of life-time income. The strength of these methods is that non-linearities and non-normalities, that may be important attributes of the earnings process, can be more easily uncovered by avoiding strong parametric assumptions.

An early important paper in this strand of the literature is Guvenen, Ozkan, and Song [2014] that documents that the distribution of earnings changes is negatively skewed and becomes more so during recessions. Thus, during recessions large upward earnings movements become less likely, while large drops in earnings become more likely. Guvenen et al. [2021] go further by analyzing life-cycle variation in skewness and other properties of the distribution of earnings changes, such as kurtosis, and how they vary with earnings levels and age. Relative

to these papers, our contribution is to analyze the same issues in the context of the Mexican labor market, a market that significantly differs from the American labor market studied by those authors and the labor markets in more advanced economies. Unlike the United States and other developed countries, the labor market experience of Mexican workers is heavily shaped by the lack of a strong social safety net, such as unemployment insurance, and the prevalence of informality —both in terms of informal jobs at formal firms and informal jobs at informal firms.

Our work is also related to Binelli and Attanasio [2010] who analyze the level and dispersion of earnings using data from Mexican household surveys for the period 1987–2002. These authors document a significant increase in wage and income inequality (P90–P10 dispersion) during the first half of the 1990s. For hourly wages, they find that the increase was characterized by inequality growing faster at the top (P90–P50) than at the bottom (P50–P10) of the distribution. For the second half of the 90s, they observe a drop or stable trend, with the slowdown in income and wage inequality being explained primarily by the top of the distribution —top-tail dispersion (P90-P50) of hourly wages and household income dropped significantly, while bottom-tail dispersion (P50–P10) decreased slightly for wages and maintained an upward trend for income. Relative to these authors, we examine a later period in the Mexican economy (2005 to 2019) and base our analysis on social security records, instead of household surveys, that are particularly suited to study earnings dynamics —our primary focus is on the distribution of earnings changes rather than on the distribution of earnings levels. Other contributions to the study of income inequality in the Mexican context are Esquivel, Lustig, and Scott [2010], Lustig, Lopez-Calva, and Ortiz-Juarez [2013], and Campos Vázquez and Lustig [2019]. The first two find that inequality decreased in Mexico from the mid 1990s to 2006, mainly due to equalizing changes in the distribution of labor income imputable to the skill premium —measured as the gap between the wages of workers with tertiary education (or secondary) and workers with no schooling or incomplete primary school— falling systematically. The third one argues that issues such as non-response and under-representation of high-wage earners and weights assignment in survey data, are relevant and can lead to mixed results in terms of evaluating the evolution of inequality. In contrast with these authors, we quantify inequality using exclusively administrative data that, while lacking potentially valuable information on education, do not suffer from the same, possibly severe, problems of survey data.

Finally, some of our findings illustrate how trying to answer the same question using different data sources may lead to different results, a pattern that has been well documented in the literature (see, for example, Celik, Juhn, McCue, and Thompson [2012], Armour,

Burkhauser, and Larrimore [2013], or Abowd and Stinson [2013]). This highlights that conclusions regarding the evolution and dynamics of earnings can be sensitive to the choice of information sources and that not every source is equally adequate to answer a specific research question.

2 Data and Macroeconomic Context

2.1 Overview of the Data

The core statistics presented in the next section of the paper are based on social security records from the Instituto Mexicano del Seguro Social (IMSS), one of the main Mexican social security institutions. All formal private sector workers who receive a salary are required, by law, to register with IMSS. The set of workers affiliated with IMSS represents approximately 80% of formal sector workers with access to social security, according to estimates from the Secretaría de Trabajo y Previsión Social (the Mexican Ministry of Labor), but it includes only a tiny fraction of government workers and does not include workers employed in the informal sector. Since informality is prevalent in Mexico, a large portion of the labor force is not included in the social security data. Own-account workers —individuals who work on their own and without employees— can register with IMSS to obtain access to some parts of the social security system and hence may appear in the social security records. By default, they are recorded with a wage equal to the minimum wage. For any given month, the share of enrolled workers that are "own-account" is roughly 0.1% of the total observations.

The social security data have a monthly frequency for the period January 2005 to December 2019, and cover, approximately, between 13 million workers at the start of the sample and 20 million workers toward the end. For the purposes of the first half of the analysis, the key variable contained in the social security data is the information on wages, reported as a worker's daily taxable income ("salario base de cotización"). This means that the data on daily wages can include various forms of compensation received by the worker other than wages (usually paid vacation and end of the year bonus), but may exclude others (in general any additional benefit or compensation that is not subject to labor income taxation), hence not necessarily reflecting the total labor income a worker receives. The data is bottom and top coded. At the bottom, the threshold is set by the minimum wage, while at the top the cap is set at 25 minimum wages before February 2017, and 25 UMAs (unit of measure and update) afterward.[3] On average, only 1.3% of the observations are bottom coded and 1.7%

[3]It is important to note that the top code is a manipulation of the data because, even if salaries above 25

are top coded. There is, however, no information on the number of hours/days/etc. worked and on whether employment was part time or full time. An important issue pertaining to the measurement of a worker's labor income is that, in a given month, the data may contain multiple observations for the same social security number (SSN), corresponding to different jobs held by the same worker (possibly with different employers). There can also be multiple observations for pairs of SSN and employer id ("registro patronal"), for which the value of the wage variable does not coincide. Multiple SSNs, nevertheless, represent no more than 2% of the observations and affect, on average, just 1.4% of the original records. In the absence of information on hours worked in each job and to avoid overstating earnings, we consider only the highest wage for those few workers who report more than one salary.

Since the information on wages is available as daily wage with monthly frequency, we first express daily wages as monthly wages multiplying the "salario base de cotización" by 30 and then add the monthly wages up to obtain the annual labor income for each worker in a given year. For the period 2005–2019, this results in over 315 million worker-year pairs with observations per year ranging between 17 and 26 millions for workers aged 14–75 years old. This constitutes our universe of potential observations. By imposing two "admissibility" conditions on the population, we construct the master sample for the analysis carried out in sections 3.1–3.3. In particular, we impose that: (i) individuals must be between 25 and 55 years of age (i.e. the prime-age labor force), and (ii) individuals must display *meaningful attachment to the labor force*, in the sense that their earnings must be above a minimum earnings threshold $Y_{min,t}$. Since in Mexico the minimum wage is defined as a daily wage, rather than as an hourly wage as it is common in other countries, we set $Y_{min,t}$ equal to 45 days of minimum wage, that corresponds to half a quarter of full-time minimum wage employment.[4] Within the subset of the sample that satisfies the first admissibility condition, the fraction of observations that are above the minimum earnings threshold varies between 97.5 and 98.5% throughout the sample period.

Table 1 presents some basic descriptive statistics and demographic characteristics in selected years for the cross-sectional sample used to carry out the analysis presented in

minimum wages/UMAs are observed in the data, the actual data are modified to record the top code in these instances. Conversely, the bottom code is not the result of data manipulation and it should be considered as actual data because, by law, no formal worker can be paid below the minimum wage and hence no worker receiving less than the minimum wage can be observed in the IMSS database. This means that the bottom code cannot affect in any way the results while, as exhaustively mentioned in the paper, the top coding requires further steps (e.g. imputation) to be dealt with.

[4]Note that for most other countries in the Global Income Dynamics Project, the threshold $Y_{min,t}$ is defined as 13 weeks of part time (20 hours a week) minimum wage employment. We chose the threshold for Mexico to be as close as possible to this common definition.

Table 1: DESCRIPTIVE STATISTICS FOR SELECTED CROSS-SECTIONAL SAMPLES

Year	Obs.	Mean income		Women	Age Shares %		
	(millions)	Men	Women	% share	[25–35]	[36–45]	[46–55]
2005	12.43	6,114	5,006	34.75	52.41	31.24	16.35
2019	19.58	6,543	5,564	38.70	46.03	31.64	22.33

Year	P5	P10	P25	P50	P75	P90	P95	P99
2005	401.89	682.54	1,584.90	3,191.88	6,483.40	12,492.76	19,602.04	53,289.61
2019	523.28	850.17	1,939.84	3,442.34	7,093.79	13,490.63	20,679.88	53,427.10

Note: Based on authors' calculations with IMSS data. The table shows summary statistics and demographic characteristics in selected years for the cross-sectional sample used to carry out the analysis presented in sections 3.1–3.3. The mean and percentiles of the income distribution are calculated using raw real earnings deflated with the Mexican CPI for 2018 and then converted to US dollars to facilitate the comparison across countries. Since the data are top coded, the percentiles above P95 are imputed by fitting a Pareto distribution around the top code. The Mexican administrative data do not contain information about the educational level of the worker.

sections 3.1–3.3. Additional details regarding the IMSS data and some relevant summary statistics for the master sample are provided in appendix A.

2.2 Macroeconomic Context in Mexico

Before moving to the core analysis, we provide a brief overview of the macroeconomic context in Mexico during the period covered by our administrative data, 2005–2019. This overview is intended to offer a minimal relevant background that may aid in the interpretation of the inequality and income dynamics results presented in the upcoming sections.

The first panel of figure 1 presents the evolution of real GDP growth. Unsurprisingly, there is a brief but noticeable contraction in GDP during the period of the 2008–2009 financial crisis. After the recovery from that recession, we observe a slowdown in growth during 2013 and a moderate contraction in 2019. Outside of those years, the 2005–2019 period can be overall characterized as one of weak, but steady growth. The second panel of figure 1 presents a Macroeconomic Uncertainty Index (MUI) for Mexico derived using the methodology outlined in Jurado, Ludvigson, and Ng [2015].[5] The economic slowdown during the financial crisis is paired with increased levels of uncertainty, which also steadily increased from mid-2016 onward. This recent trend can be associated with both external and internal factors. On the external side, the renegotiation of NAFTA, along with a generalized more protectionist stance from the United States, represented a large and significant increase in trade policy

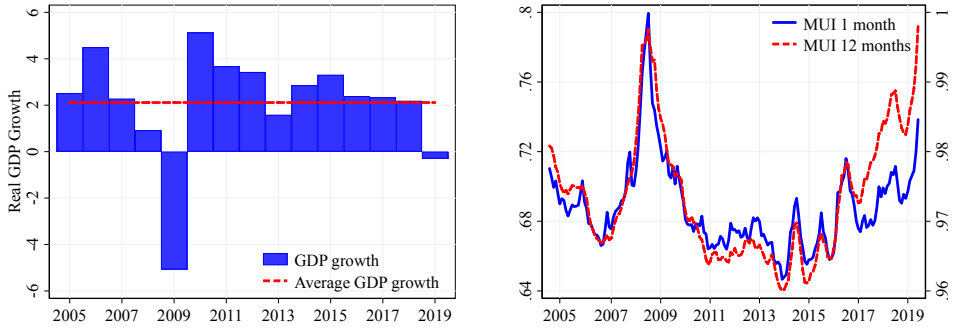
[5]In the presence of adjustment costs in both labor and capital inputs, uncertainty can have a significant impact on firm's investment, job creation/destruction decisions and, hence, on labor market outcomes. For example, Bloom [2009], Jurado et al. [2015], and Baker, Bloom, and Davis [2016] have shown that increases in uncertainty can be associated with reductions in employment and investment, while Mathy [2020] finds evidence suggesting that an increase in uncertainty can negatively affect wages.

uncertainty for Mexico vis-a-vis its largest trading partner. This source of uncertainty did not fully subside until early 2020 with the enactment of the USMCA. On the internal side, a period of increasing uncertainty commenced with the presidential election and the sudden cancellation of the new Mexico City airport project in the second half of 2018.

Figure 1: AGGREGATE ACTIVITY AND ECONOMIC UNCERTAINTY

(a) Real GDP growth

(b) Macroeconomic Uncertainty Index (MUI)



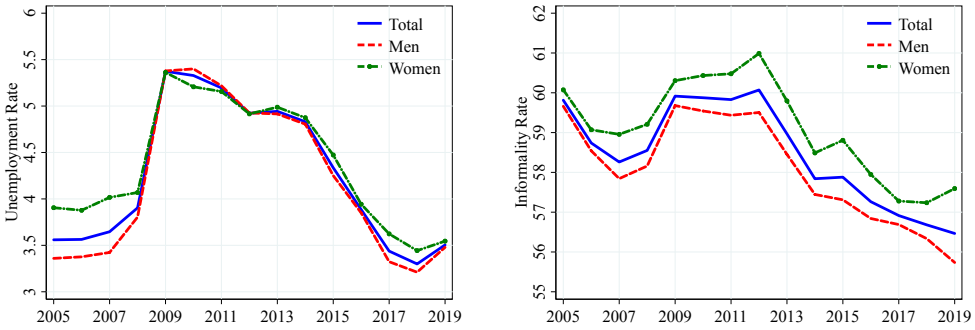
Note: Based on authors' calculations with data from INEGI (the Mexican statistical agency). In panel (a), the red dashed line refers to the average GDP growth during the sample period. In panel (b) the MUI is calculated based on information from 125 monthly economic series using the methodology outlined in Jurado et al. [2015]. The MUI 1-month (left axis) refers to the uncertainty index constructed based on one-month ahead forecast errors and the MUI 12-months (right axis) refers to the index constructed based on 12-month ahead forecast errors.

Regarding aggregate labor market conditions, figure 2 presents the unemployment and informality rates. The informality rate is the percentage of the employed population (15 years and older) that is employed in an informal job. In Mexico, a job is classified as informal if it is performed without legal or institutional protection, regardless of the nature of the economic unit/firm in which said job is carried out, while a worker is classified as informal if he/she is a wage earner who does not have access to social security and/or a own-account worker who does not follow a formal accounting system. The unemployment rate increased during the financial crisis, reaching 5.5% in 2009. This number is low compared to the United States, where it increased to 10% in the same period. Between 2012 and 2014, unemployment remained stable before resuming its decreasing trajectory toward pre-financial crisis levels. By the end of the period under consideration, it started to moderately increase. Differentiating by gender, we see that the unemployment rate is fairly similar for men and women, with the exception of the 2005–2008 and 2016–2018 periods when unemployment for women was noticeably higher than for men. The relatively low unemployment rate in Mexico, even in the

middle of a crisis of significant magnitude, is consistent with workers being unable to stay unemployed due to the lack of a social safety net, particularly unemployment insurance. The rate of informality is quite high, ranging between 56 and 60% during the entire period with the exception of the years 2008–2012 when it displayed a steady downward trend. Informality is higher for women than for men. Over the period considered, there is a strong correlation between the rate of informality for men and women, except in 2017–2019 when the rate of informality for men kept trending downward, while that for women showed a moderate increase.

Figure 2: UNEMPLOYMENT AND INFORMALITY

(a) Unemployment rate **(b)** Informality rate

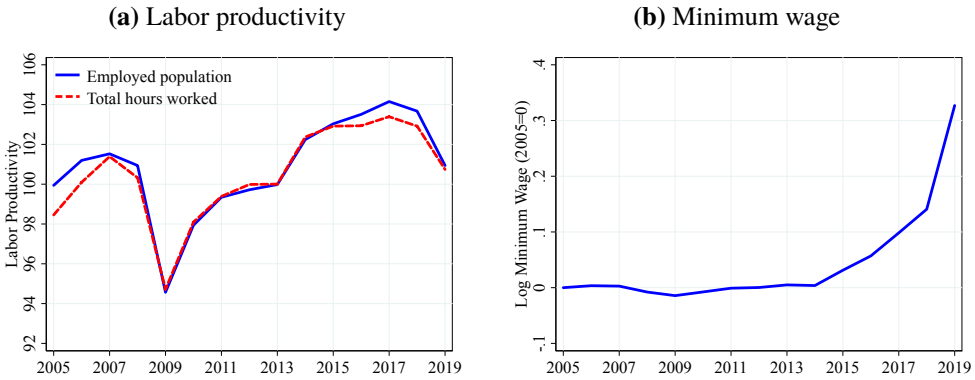


Note: The unemployment rate in year t corresponds to the average quarterly unemployment rate in that year. The rate of informality for year t corresponds to the average monthly rate of informality throughout that year. The official statistics published by INEGI are the quarterly and monthly rate, respectively.

The first panel of figure 3 shows the evolution of labor productivity, based on the employed population and the total hours worked. During the financial crisis there was a sharp drop in labor productivity, and pre-crisis levels were not reached again until 2014. Starting in 2013 there was a strong increase that peaked in 2017. From 2017 labor productivity started trending downward. The sharp increase in labor productivity between 2013 and 2017 coincides with a broad labor market reform that was approved at the end of 2012 and addressed a variety of issues such as outsourcing, conditions for trial hires and hires with flexible schedules, paternity leave, union transparency, regulation of under-age work, among others. The right panel in figure 3 illustrates the evolution of the (real) minimum wage that remained largely unchanged for most of the period of analysis. Starting in 2014 the minimum wage has been increasing steadily and substantially. The evolution of the minimum wage will be particularly useful for understanding the time-series patterns of earnings for the lowest percentiles of the

income distribution, since the social security data are bottom coded at the minimum wage.[6] Remarkably, the sharpest increases in the minimum wage have occurred from 2017 onward, a period that also coincides with the decrease observed in aggregate labor productivity.

Figure 3: LABOR MARKET CONDITIONS



Note: Based on authors' calculations with data from INEGI and IMSS. Labor productivity in year t corresponds to the average, seasonally adjusted, quarterly labor productivity index for all the quarters of year t. The labor productivity index is calculated as the ratio between value of production and hours worked or employed workers at constant prices. The minimum wage, in logs, is annualized, deflated with the 2018 Mexican CPI and normalized to 0 in 2005.

3 Core Statistics on Inequality, Mobility and Income Dynamics: Evidence from Social Security Data

In this section we present a descriptive characterization of inequality, earnings dynamics, and mobility based on Mexican administrative records. The non-parametric approach used in our analysis is closely related to the work of Bonhomme and Hospido [2017] for Spain and Guvenen et al. [2017] and Guvenen et al. [2021] for the United States. To the best of our knowledge, we are the first to provide such a characterization of the earnings distribution of private sector formal workers in Mexico that is based on administrative records and, together with Engbom, Gonzaga, Moser, and Olivieri [2022] and Blanco, Díaz de Astarloa, Drenik, Moser, and Trupkin [2022], the first to do so in the context of an emerging economy whose labor market significantly differs from those of more advanced economies.[7] Note that all

[6]Recall that, on average, 1.3% of the observations in IMSS are bottom coded, which implies that (at least) 1.3% of the formal workers are those who should have directly benefitted from the minimum wage increases.

[7]Binelli and Attanasio [2010] present some results that are related to the analysis presented in this section and in appendix B, particularly concerning income inequality. These authors base their analysis on Mexican

the results presented in sections 3.1–3.3 distinguish between men and women, results for the whole sample are presented in appendix B.

3.1 Income Inequality

We begin our descriptive analysis by characterizing the most salient properties of the distribution of (log) real earnings. Figure 4 presents the evolution of the percentiles (relative to 2005) of the earnings distribution for both men and women.[8] Notably, the 2008–2009 financial crisis had a negative impact on earnings across the whole earnings distribution, but especially so for men and for workers at the bottom of the distribution. The earnings of workers at the very top proved to be more resilient to the negative shock associated with the crisis. Between 2013 and 2014 there was a decrease in the earnings of the lowest percentiles, particularly sharp for men. This is likely associated with the fact that 2013 was a year of below-average GDP growth and stalled growth in labor productivity (see figures 1 and 3). Outside of these two periods, real earnings have shown an upward trend that is more noticeable for women than for men. Male workers showed more ups and downs at the bottom of the earnings distribution and flatter profiles at the top. In contrast, female workers displayed a steadier upward trend across the entire earnings distribution.

Starting in 2014 for men and in 2016 for women, the lowest percentiles of the earnings distribution (P10 and P25) display a significant upward trend relative to previous years with log earnings increasing by roughly 20 log points between 2016 and 2019. This pattern appears to be highly correlated with the evolution of the minimum wage (see right panel of figure 3). The fact that the evolution over time of the lowest percentiles of the log earnings distribution correlates with the evolution of the minimum wage is not particularly surprising since the administrative data are bottom coded at the minimum wage. It is, however, noteworthy to observe that this correlation can be observed through P25, suggesting that increases in the minimum wage may have spillover effects on wages higher up in the earnings distribution. This result, known in the literature as the "lighthouse effect", echoes the findings of Engbom and Moser [2021] for Brazil and Kaplan and Novaro [2006] and Campos Vázquez and Rodas Milián [2020] for Mexico. The increase in earnings in low percentiles that are not

household surveys, rather than administrative data, and their results are focused on the period 1987 to 2002, a period that is not covered by our administrative records.

[8]When looking at figure 4 it is important to note that, as the administrative data are top coded, only the percentiles up to P95 are derived using actual data. The percentiles after P95 are imputed by fitting a Pareto tail distribution around the top code. We include these upper percentiles for completeness and for the sake of comparison with other countries even though the information provided after P95 is inevitably noisy and should be interpreted with caution.

directly affected by the minimum wage may also relate to increases in labor productivity after 2013. Even though labor productivity declines slightly toward the very end of our sample period, its levels in 2018 and 2019 are still above 2013's levels.[9]

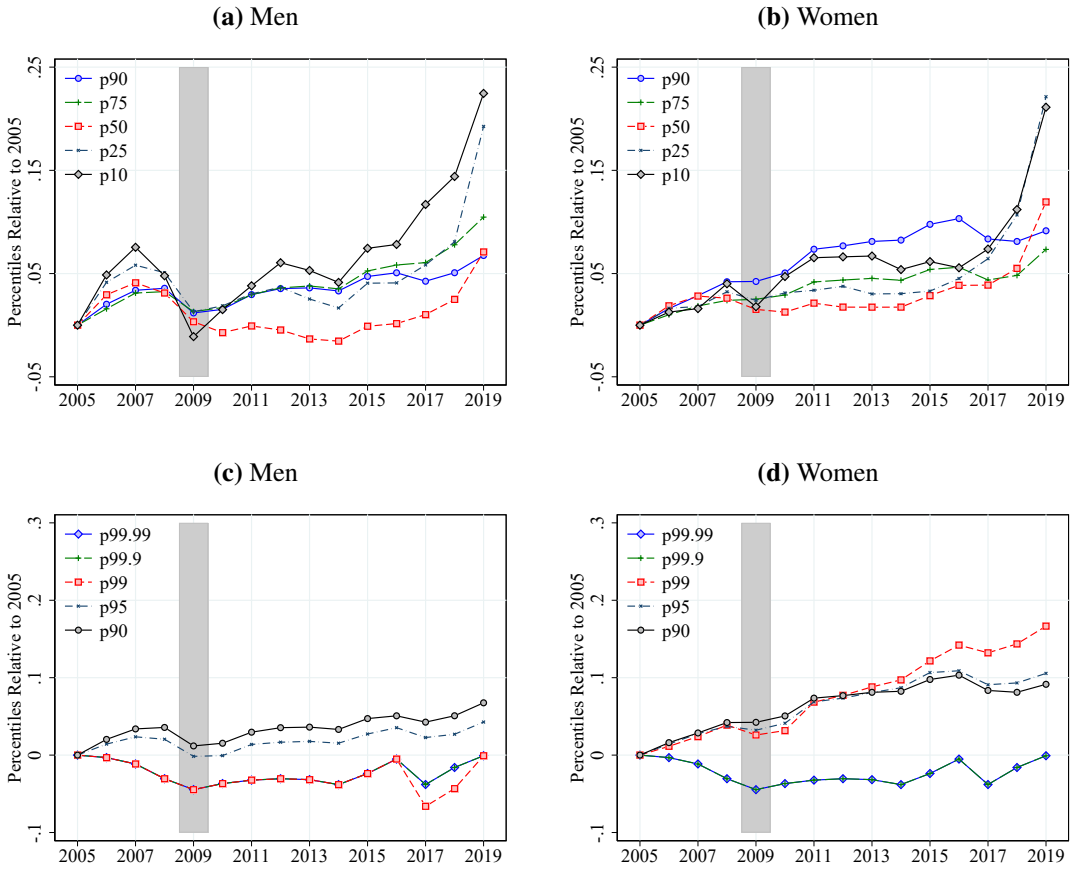
The evolution of several measures of dispersion in the distribution of real earnings is documented in figure 5. Panels (a) and (b) show that the overall dispersion in this distribution is very similar for men and women and seems to closely match the dispersion that would be observed in a normal distribution. The dispersion is, on average, slightly higher for men, who are also those who experienced a more noticeable increase in earnings dispersion during the crisis of 2008–2009. Outside of this recessionary period, the dispersion remained fairly stable between 2005 and 2015, but started to display a steady decreasing trend from 2015 onward.[10] The behavior of the P90-P10 measure presented in panels (a) and (b) can be further understood by looking at the the evolution of the upper and lower tails of the distribution —P90-P50 and P50-P10, respectively— presented in panels (c) and (d). The relative stability of P90-P10 from 2005 through 2015 is the result of a slight downward trend in the lower tail of the distribution that is barely offset by a slight upward trend in the upper tail. The downward trend of P90-P10 from 2015 onward can be associated with a downward trend in both the lower (P50-P10) and upper (P90-P50) tails of the distribution. Referring back to figure 4, the reduction in the inequality of log earnings is mainly driven by the growth of the lower percentiles paired with the relative stability in the highest percentiles.

Finally, figures 6 and 7 describe how initial inequality, for workers aged 25, has evolved over time and how life-cycle inequality differs across different cohorts of workers. In figure 6 we see that the lower tail (P50-P10) experience a modest downward trend, while the upper tail (P90-P50) experience a very slight upward trend. These trends are similar for men and

[9]Historically, there have been several additional economic factors related to changes in wages in Latin America and, in particular, in Mexico. Trade and globalization have been identified as relevant to the Mexican economy, given that NAFTA integration in the 1990s reduced unemployment, and boosted employment and wages, while Chinese competition tended to have the opposite effect in the 2000s (Chiquiar, Covarrubias, and Salcedo Cisneros [2017], Blyde, Busso, and Romero [2020]). Increases in education of the workforce was also relevant to rise low-skilled wages during the 2000s and to reduce skill premia (Jaume [2021]), consistent with the race between education and technology discussed in the literature (Katz and Murphy [1992], Goldin and Katz [2010]). The role of education, however, has been less pronounced in Mexico due to a lower educational upgrading with respect to other Latin American countries (Lustig et al. [2013], Acosta, Cruces, Galiani, and Gasparini [2019]). To disentangle the relative importance of each of these factors in contributing to the increases in the lowest percentiles of the earnings distribution that we document in this paper constitutes an important research agenda but is beyond the scope of this paper.

[10]Figures B.4 and B.5 in appendix B are also consistent with decreasing earnings inequality from 2016 onward. In particular, B.4 shows that the share of income that accrues to the bottom 50% of the earnings distribution grew during that period, while the income share of the top 10% of the distribution decreased. Similarly, B.5 shows that the Gini coefficient of the earnings distribution decreased substantially between 2016 and 2019, relative to the trend it had displayed in previous years.

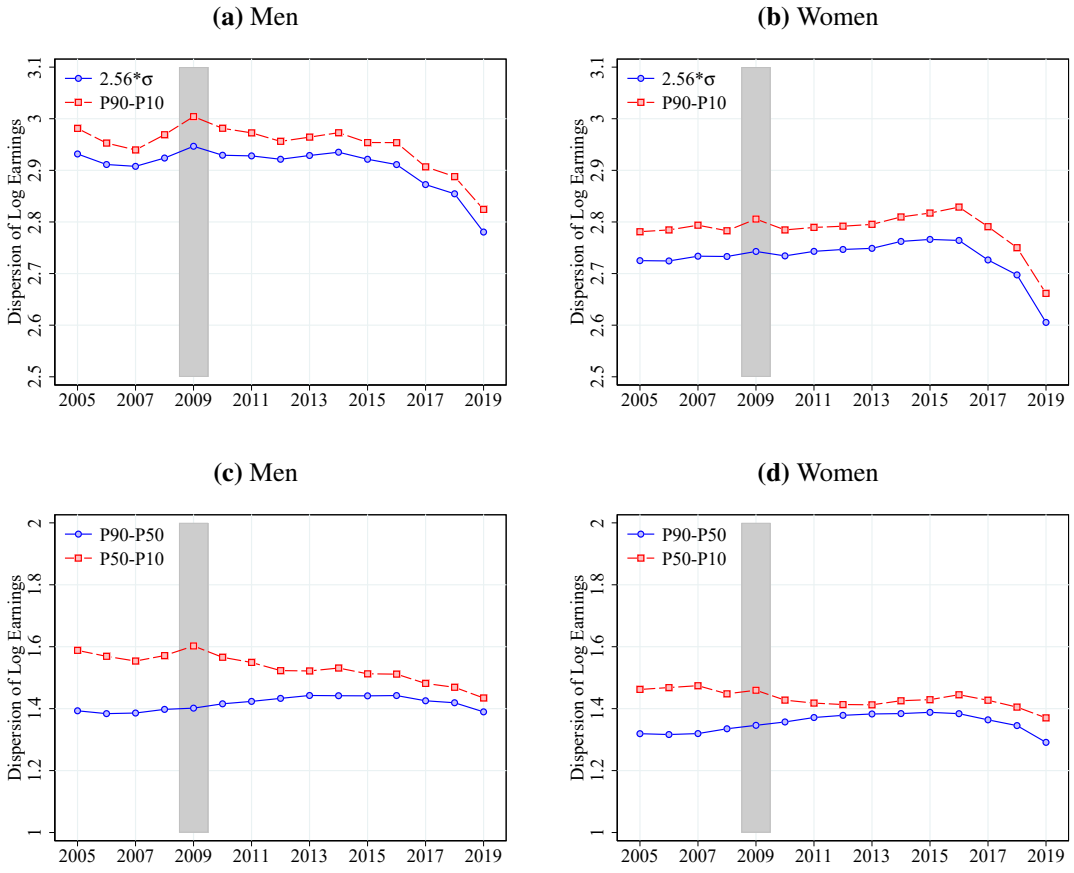
Figure 4: EVOLUTION OF THE PERCENTILES OF THE LOG REAL EARNINGS DISTRIBUTION



Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time the following percentiles of the distribution of log (real) earnings: (a) Men: P10, P25, P50, P75, P90; (b) Women: P10, P25, P50, P75, P90; (c) Men: P90, P95, P99, P99.9, P99.99; (d) Women: P90, P95, P99, P99.9, P99.99. Since the data are top coded the percentiles above P95 are imputed by fitting a Pareto distribution around the top code. All percentiles are normalized to 0 in 2005, the first available year. Shaded areas are recessions.

women, but are marginal as new cohorts of young workers faced relatively stable earnings dispersion throughout our sample period. The patterns of life-cycle inequality across different cohorts are also very similar for men and women as illustrated in figure 7. Specifically, for any given cohort, dispersion of log earnings increases until the last years of the sample (until 2016 or 2018 depending on the cohort) and displays a downward trend thereafter. Once again, this is likely associated with the increase in earnings for minimum wage workers starting around 2016 and the relative stability of the real earnings for workers at the top of the earnings distribution. Across cohorts, the dispersion of earnings for workers aged 25 shows a very slight downward trend, except for those being 25 around the time of the 2008–2009 crisis

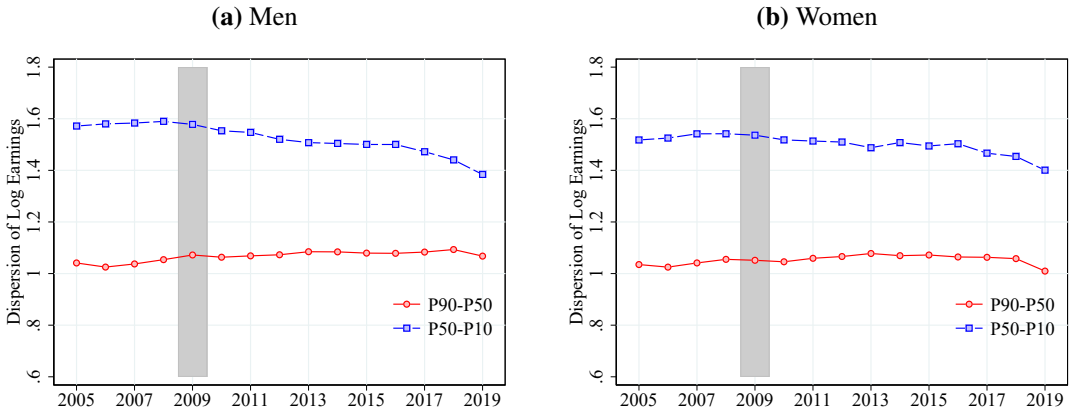
Figure 5: EVOLUTION OF THE PERCENTILES OF THE LOG REAL EARNINGS DISTRIBUTION



Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time the following measures of overall, top- and bottom-tail dispersion of the distribution of log earnings: (a) Men: P90–P10 and 2.56*σ (sigma is the standard deviation); (b) Women: P90–P10 and 2.56*σ of log income; (c) Men: P90–P50 and P50–P10; (d) Women: P90–50 and P50–P10. Shaded areas are recessions. 2.56*σ corresponds to the P90–P10 differential for a Gaussian distribution.

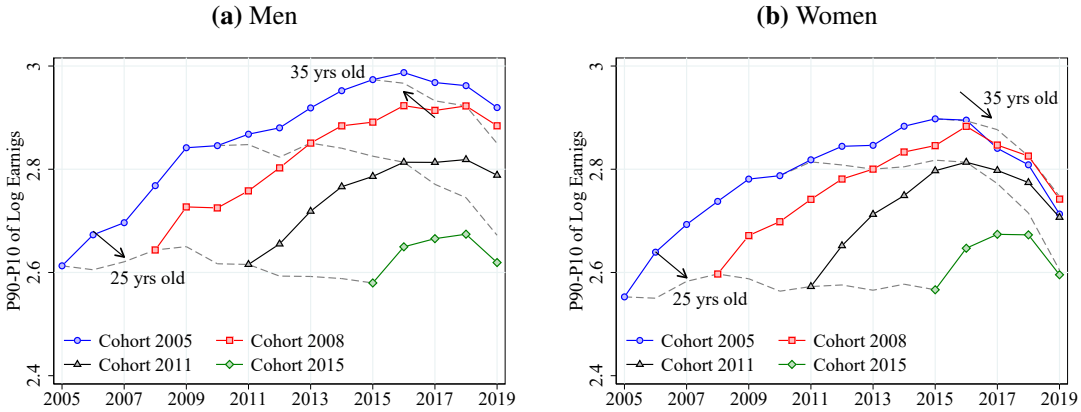
when dispersion increased. In contrast, there is a more noticeable downward trend in the dispersion of earnings across cohorts for workers 30 and 35 years of age. Given the relatively short window covered by our social security records, it is however unclear to what extent the eventual decrease in dispersion documented in figure 7 can be attributed to features of a worker's life-cycle as opposed to the overall decrease in dispersion observed in the latter part of our sample period.

Figure 6: INITIAL INEQUALITY: DISPERSION OF LOG EARNINGS FOR WORKERS AT AGE 25



Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time the following measures of top- and bottom-tail dispersion of the log earnings distribution: (a) Men: P90–P50 and P50–P10 for workers at age 25; (b) Women: P90–P50 and P50–P10 for workers at age 25. Shaded areas are recessions.

Figure 7: LIFE-CYCLE INEQUALITY ACROSS COHORTS



Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time measures of overall dispersion of the log earnings distribution: (a) Men: P90–10 over the life cycle for the 2005, 2008, 2011, and 2015 cohorts; (b) Women: P90–10 over the life cycle for the 2005, 2008, 2011, and 2015 cohorts. The grey dashed lines link across cohorts the years corresponding to ages 25, 30, and 35.

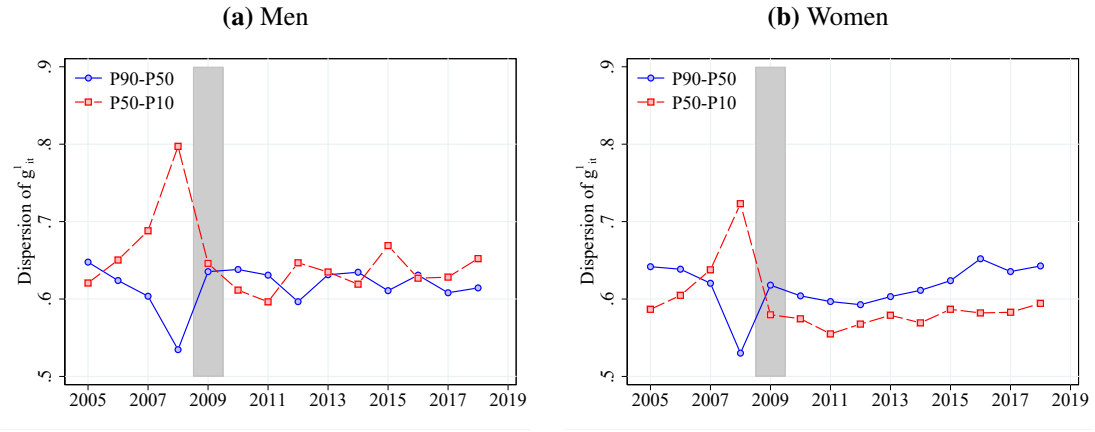
3.2 Income Dynamics

The results of this section constitute the core of our analysis of the administrative data. We characterize the most relevant properties of individual earnings dynamics, with an emphasis on the statistics that provide a diagnostic of the extent to which the distribution of earnings

changes deviates from normality. As in Guvenen et al. [2021], the approach is non-parametric to avoid the strong assumptions embedded in benchmark econometric models of earnings dynamics, as those could mask important features of the distribution of earnings changes. Deviations from normality are important to study because they have direct implications for the kind of income shocks a worker may experience, such as their magnitude, direction (higher likelihood of positive vs. negative shocks), and frequency of extreme events.

Defining (residualized) log earnings changes as $g_{it}^{k} = \Delta^{k}\varepsilon_{it} = \varepsilon_{i,t+k} - \varepsilon_{it}$, with ε_{it} being the residuals of a regression of log earnings on a full set of age dummies, separately by gender and year, Guvenen et al. [2021] show that as k becomes larger, the distribution of earnings changes $\Delta^{k}\varepsilon_{it}$ reflects more closely the distribution of the permanent component of earnings changes rather than that of transitory innovations. Here we focus on one-year earnings changes (i.e. $k = 1$), that reflect mainly transitory innovations to earnings, and present results for $k = 5$, the more permanent innovations, in appendix B.

Figure 8: Dispersion of one-year log earnings changes



Note: Based on authors' calculations with data from IMSS. Using the time-series dimension of the sample, this figure plots against time the following measures of top- and bottom-tail dispersion of the distribution of one-year earnings changes: (a) Men: P90–P50 and P50–P10 differentials; (b) Women: P90–P50 and P50–P10 differentials. Shaded areas are recessions.

Figure 8 illustrates the dispersion in the upper (P90–P50) and lower (P50–P10) tails of the distribution of one-year log earnings changes. For both men and women, we see that between 2008 and 2009, that is between the onset and bottoming out of the depressed economic conditions caused by the financial crisis, there was an increase in dispersion in the lower tail and a decrease in the upper tail of the distribution. The magnitude of these opposing movements suggests that overall dispersion (P90–P10) increased for men but remained relatively stable for

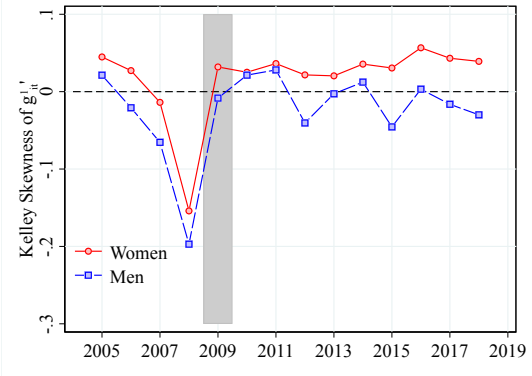
women during the time of said crisis. It is worth noting that these spikes in dispersion coincide with the sharp increase in unemployment and informality that occurred between 2008 and 2009 (see figure 2). Outside of this recessionary period, we observe only modest movements in P90–P50 and P50–P10. These two measures of dispersion move up and down in opposing directions for men, albeit the movements are small so that P90–P10 remains relatively stable. In contrast, from 2011 onward women experience a slight upward trend in both P90–P50 and P50–P10 resulting in a greater overall dispersion in their distribution of one-year earnings changes. This period of increasing dispersion for women coincides with a post-labor market reform period in which female labor force participation grew in Mexico, likely due, at least in part, to the more flexible conditions encouraged by the reform.[11] Thus, it is possible that the greater flexibility in hiring practices induced the entry of women into the labor market, but particularly so for those who may demand more flexible work arrangements.

For the purpose of detecting significant deviations from normality, figure 9 plots higher order moments of the distribution of one-year earnings changes. In particular, we consider Kelley skewness and excess Crow-Siddiqui kurtosis and whenever we mention skewness and/or kurtosis in this section we are referring to these two measures. Notice that, other than the particularly sharp decrease and rebound of skewness during the period of the financial crisis observed across genders, skewness remains fairly stable and relatively close to zero for both men and women from 2009 onward. Regarding kurtosis, we find that the distribution is leptokurtic, implying that, while most workers experience transitory earnings changes of a small magnitude, a non-negligible mass of them experiences extreme shocks. Qualitatively, the evolution of kurtosis is fairly similar for both men and women, although its level is significantly higher for women. Kurtosis had been trending slightly upward before the financial crisis, but increased significantly during this recession, particularly so for women. From 2011 onward, nonetheless, we see both a strong downward trend and a significant reduction in the gap level between men and women. Despite this significant downward trend in the kurtosis of one-year earnings changes, by the end of our period of study its level remains quite high. Note that, even if the levels of excess kurtosis of one-year earnings changes in the Mexican administrative data are high, they are not out of line with what Guvenen et al. [2021] report for the United States, particularly those that prevail in the latter part of our sample period. While our results suggest an average kurtosis for one-year earnings changes that is

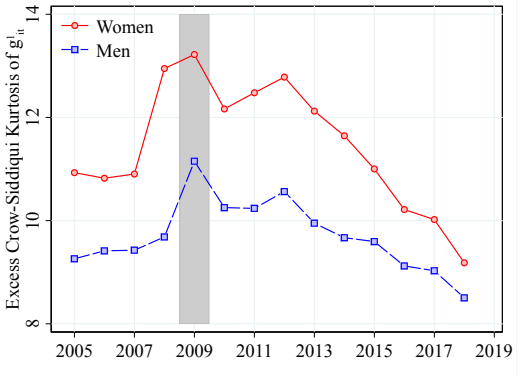
[11] De la Cruz Toledo [2018] finds that the higher preschool enrollment associated with the rollout of a universal preschool policy led to an increase in female labor force participation in Mexico. Growing female labor force participation is also reflected in our administrative data with the share of women in the IMSS master sample increasing from 36.4 to 38.7% between 2011 and 2019.

Figure 9: SKEWNESS AND KURTOSIS OF ONE-YEAR LOG EARNINGS CHANGES

(a) Kelley skewness



(b) Excess Crow-Siddiqui kurtosis



Note: Based on authors' calculations with data from IMSS. Using the time-series dimension of the sample, this figure plots against time the following higher order moments of the distribution of one-year earnings changes: (a) Men and Women: Kelley skewness calculated as $\frac{(P_{90}-P_{50})-(P_{50}-P_{10})}{P_{90}-P_{10}}$; (b) Men and Women: Excess Crow-Siddiqui kurtosis calculated as $\frac{P_{97.5}-P_{2.5}}{P_{75}-P_{25}} - 2.91$, where the first term is the Crow-Siddiqui measure of kurtosis and 2.91 corresponds to the value of this measure for the Normal distribution. Shaded areas are recessions.

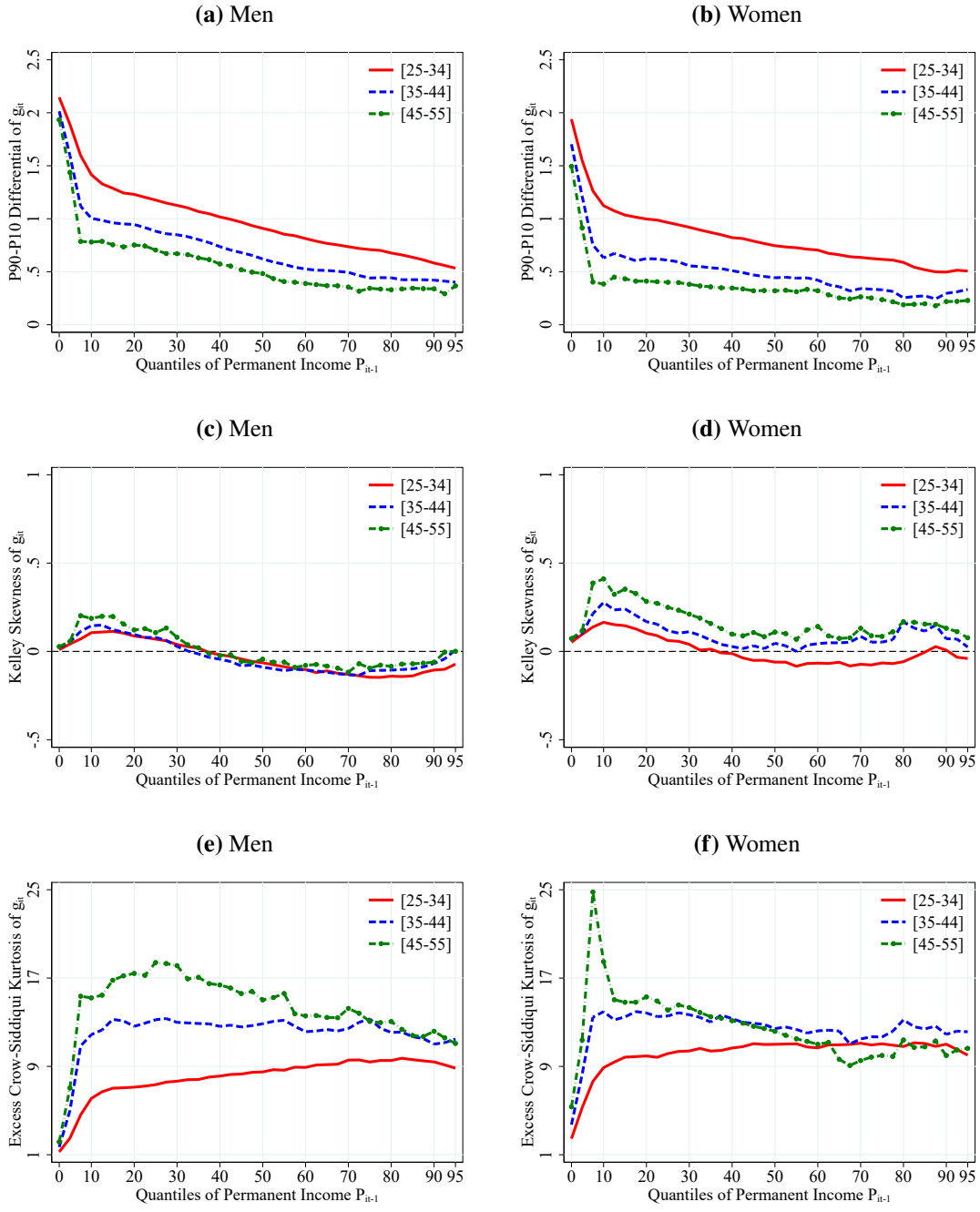
higher than the one for the United States, Guvenen et al. [2021] also find high levels of excess kurtosis of up to 11 for workers aged 45 to 54 in the 60th percentile of the distribution of recent earnings, a level that is comparable to ours for men during the whole sample period and for women from 2015 onward. Conversely, if we look at the excess kurtosis of five-year income changes (see figure B.8 in appendix B), the levels are significantly smaller than those depicted in figure 9, suggesting that the distribution of transitory earnings shocks is significantly peakier than that of permanent shocks, i.e. extreme shocks to earnings changes are more likely to be transitory than permanent.

It is also of interest to understand how the properties of the distribution of one-year earnings changes may vary along the permanent income distribution and along the life-cycle of workers.[12] To this end, figure 10 presents dispersion, skewness, and kurtosis conditional on age (for age groups 25 to 34, 35 to 44, and 45 to 55) and conditional on the percentile of a worker's permanent income. The permanent income is calculated aggregating over a period of 15 years, the maximum number of years available in our sample, from 2005 to 2019.

[12]Permanent income for each worker is defined as $P_{it-1} = \frac{\sum_{s=t-3}^{t-1} y_{is}}{3}$. This measure takes average raw earnings y_i (including zeros or earnings below the meaningful attachment to the labor force earning threshold) over the previous three years. It is constructed only for those workers who have at least two years of earnings above the threshold.

Several relevant patterns are qualitatively similar across genders, but there are also important differences. Specifically, dispersion is monotonically decreasing with both age and permanent income, meaning that transitory shocks to earnings are the most volatile for younger and lower-income workers. Skewness seems to be increasing with age. This is evident in the case of women, and for men at the bottom and top of the permanent income distribution; between P30 and P70 skewness is fairly similar across age groups. Along the distribution of permanent income, skewness decreases monotonically up to the 55th percentile, approximately, and then remains relatively stable, with a slight upward trend from P85 in the case of men. Men also display skewness that is positive for the lower part of the permanent income distribution and negative for the upper percentiles and this applies to all age groups. In contrast, for women aged 35 and older, skewness is positive along the entire distribution of permanent income. For young women, as it was the case with young men, skewness is positive at the bottom of the permanent income distribution but becomes negative from around the 40th percentile onward. The positive skewness of one-year earnings changes observed for all women 35 or older and for men of all ages in the lower part of the permanent income distribution implies that the earnings of these workers have more room to move up and less room to fall. This contrasts with the patterns observed during the financial crisis when skewness became noticeably negative, meaning that this recessionary episode, albeit brief, temporarily and significantly increased the chances for all Mexican workers of experiencing larger and more frequent negative shocks. Finally, kurtosis, is monotonically increasing with age for men. The same is true for women as well, but only up to about the 40th percentile of the permanent income distribution. For the youngest workers, kurtosis increases sharply up to the 15th percentile, it continues to increase, but very gradually, up to roughly the 85th percentile, after which it slightly trends downward. In contrast, for the oldest workers kurtosis increases up to the 25th and 8th percentile for men and women, respectively, and then trends downward. Men experience this downward trend gradually, while for women the decrease is sharp between P8 and P10 and then kurtosis starts a gradual downward trend along the rest of the permanent income distribution. Overall, older and lower-income workers face the most concentrated (peakiest) distribution of transitory income shocks.

Figure 10: DISPERSION, SKEWNESS AND KURTOSIS OF ONE-YEAR LOG EARNINGS CHANGES CONDITIONING ON PERMANENT INCOME FOR WORKERS OF DIFFERENT AGES

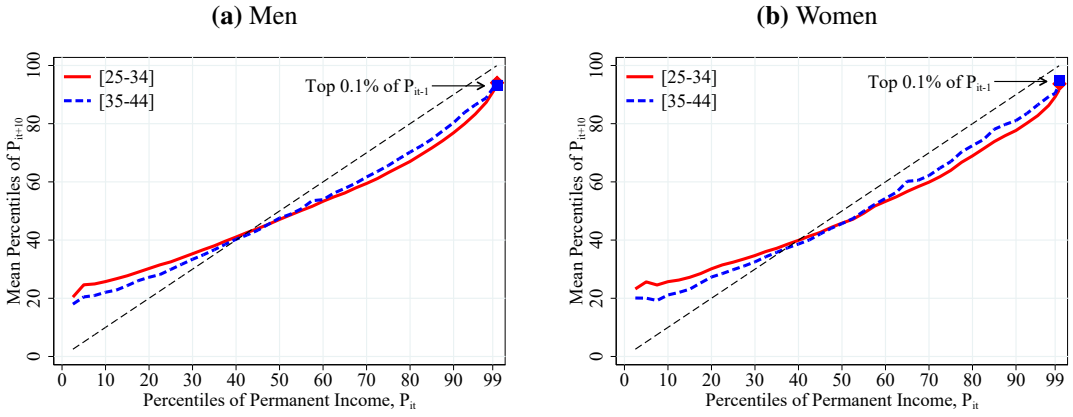


Note: Based on authors' calculations with data from IMSS. Using the worker heterogeneity dimension of the sample, this figure plots against percentiles of the permanent income distribution, and for three different age groups, the following moments of the distribution of one-year earnings changes: (a) and (b) Men and Women: P90–P10 differential; (c) and (d) Men and Women: Kelley Skewness; (e) and (f) Men and Women: Excess Crow-Siddiqui kurtosis. The permanent income is calculated aggregating over a period of 15 years, the maximum number of years available in our sample, from 2005 to 2019. Since the data are top coded the percentiles of the permanent income distribution are plotted only until P95.

3.3 Income Mobility

We now turn to the analysis of mobility over time in the distribution of earnings. Figure 11 shows long-term mobility for workers aged 25–34 and workers aged 35–44, tracking their movements along the distribution of permanent income over a 10-year time horizon. We observe upward mobility at the bottom and downward mobility at the top of the permanent income distribution, except at the very top. These patterns of mobility are qualitatively similar for men and women. For both age groups under consideration, there is upward mobility up to about P45 and P35 for men and women, respectively with upward mobility being slightly higher for younger workers. In contrast, individuals located in the top percentiles of the permanent income distribution experience downward mobility that is also higher for younger workers. At the very top of the permanent income distribution, top 0.1%, there is essentially no income mobility.

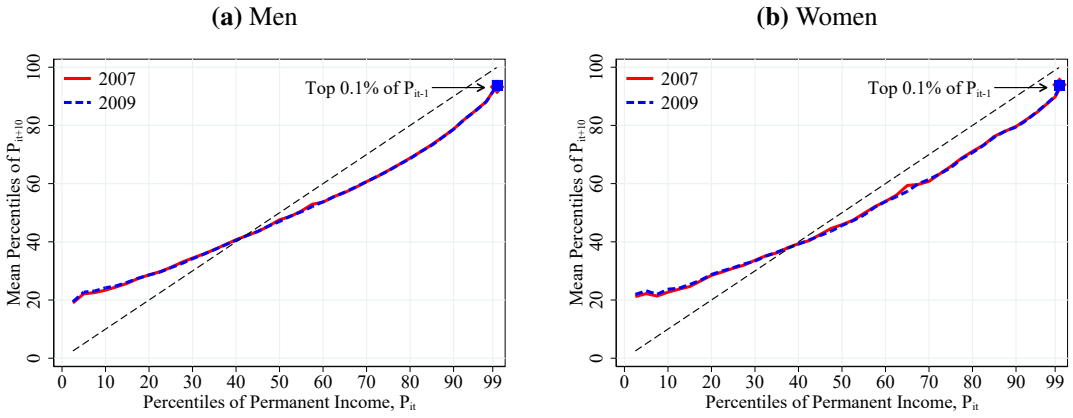
Figure 11: Evolution of 10-year mobility over the life cycle



Note: Based on authors' calculations with data from IMSS. The figure shows average rank-rank long-term (10-year) mobility for male (a) and female (b) workers of different ages.

Looking at long-term mobility over time, figure 12 shows the evolution of 10-year mobility for two different starting years, 2007 and 2009. The patterns are again very similar across men and women, but it is worth noting that 10-year mobility within the distribution of permanent income is essentially unchanged between 2007 and 2009, despite 2007 being a pre-crisis year and 2009 coinciding with the large macroeconomic shock associated with the financial crisis (see figure 1).

Figure 12: Evolution of 10-year mobility over time



Note: Based on authors' calculations with data from IMSS. The figure shows average rank-rank long-term (10-year) mobility for male (a) and female (b) workers in selected years of the sample, 2007 and 2009.

4 Administrative vs. Survey Data, Worker Transitions, and Early Exposure to Informality

An important limitation of the social security data and, hence, of the results presented in section 3 for understanding inequality and the dynamics of income in the Mexican labor market, is that these data only cover workers in the formal sector. As such, they are not informative about a large segment of the Mexican labor market: informal workers. In contrast to developed countries, a notorious characteristic of developing and emerging economies, such as Mexico, is that the informal sector is responsible for a large share of all economic activity and commands a significant proportion of the economy's productive resources. During the period considered in our analysis, 2005–2019, the quarterly rate of informality ranged between 56 and 60% at the aggregate level (see figure 2), with this rate being somewhat larger for women than for men. Furthermore, Levy [2018] shows that informal firms in Mexico represent 90% of all firms and absorb 40% of the economy's capital stock. Thus, these firms are present throughout the economy and are not confined to activities deemed "traditional" or "less modern", implying that the informal economy fully coexists alongside the formal economy.

Due to the pervasiveness of informality in the Mexican economy and the fact that the relative size of the informal sector may affect the overall functioning of the labor market and the opportunities available to workers in terms of job security, social mobility, and life-time earnings, any analysis that omits a discussion of informal employment can only provide a partial perspective regarding the labor market outcomes, including earnings, of a large share of Mexican workers. With this in mind, we devote the second part of our analysis to three exercises. First, we contrast some of the results from the previous section, based on IMSS data, with results obtained from a comparable sample from the Encuesta Nacional de Ocupación y Empleo (ENOE for its acronym in Spanish), the main household survey on employment in Mexico. In addition, we provide results for informal workers and the pooled sample of formal and informal workers. This exercise is informative about the extent to which household survey data could be used to provide a more comprehensive view of the Mexican labor market by offering information on both formal and informal workers. Second, we study the wage dynamics of workers who exit and subsequently re-enter the IMSS database, as a large number of workers in Mexico maintain a tenuous connection with formal employment. Even though we cannot track workers upon exit from the social security data, the very low levels of unemployment and absence of a social safety net in Mexico suggest that the majority

of these workers are likely to be transitioning in and out of formal employment via job spells in the informal sector. This is especially plausible for men, whose labor participation rate is above 90% for the age group considered in the analysis. Finally, we quantify the impact that having the first job in the informal sector has on future earnings.

The rest of this section is organized as follows. Section 4.1 presents the comparison between IMSS-based statistics and ENOE-based statistics. Section 4.2 analyzes the impact of transitioning out of and back into IMSS-affiliated employment on the wages earned within the formal sector. Section 4.3 presents a quantification of the effect of having the first job in the informal sector on future earnings.

4.1 Earning Inequality and Dynamics: Comparing Administrative and Survey Data

To provide a more comprehensive picture of the distributions of earnings levels and changes of Mexican workers and evaluate the evolution of income inequality for formal and informal workers, we rely on an additional source of information, the Encuesta Nacional de Ocupación y Empleo (ENOE) in this section. The ENOE is the primary official source of employment and occupation data in Mexico. It is a quarterly household survey at the national level run by the Mexican Statistical Agency (INEGI) and is specifically designed to collect information on the employment situation of individuals in rural and urban areas. It also collects information on labor income and socioeconomic characteristics of the individuals surveyed. It has a rotating panel structure: every quarter, INEGI replaces one-fifth of the sample (i.e. each household is followed for five consecutive quarters before being dropped from the survey). The ENOE is representative at the national and state level, and for one selected city in each state. It is conducted continuously throughout the year, surveying approximately 120 thousand households in each quarter.

It is well known that using survey data presents several challenges, such as measurement error, lack of representativeness in the tails of distributions, small sample size, increasing non-response, and declining response quality (see National Research Council [2013] and Meyer, Mok, and Sullivan [2015]). Despite these limitations that the ENOE also suffers from, it provides the necessary information to compute different indicators that can be compared to those constructed using the administrative data presented in section 3.1 (cross-sectional evolution of income inequality) and section 3.2 (transitory shocks to earnings).[13]

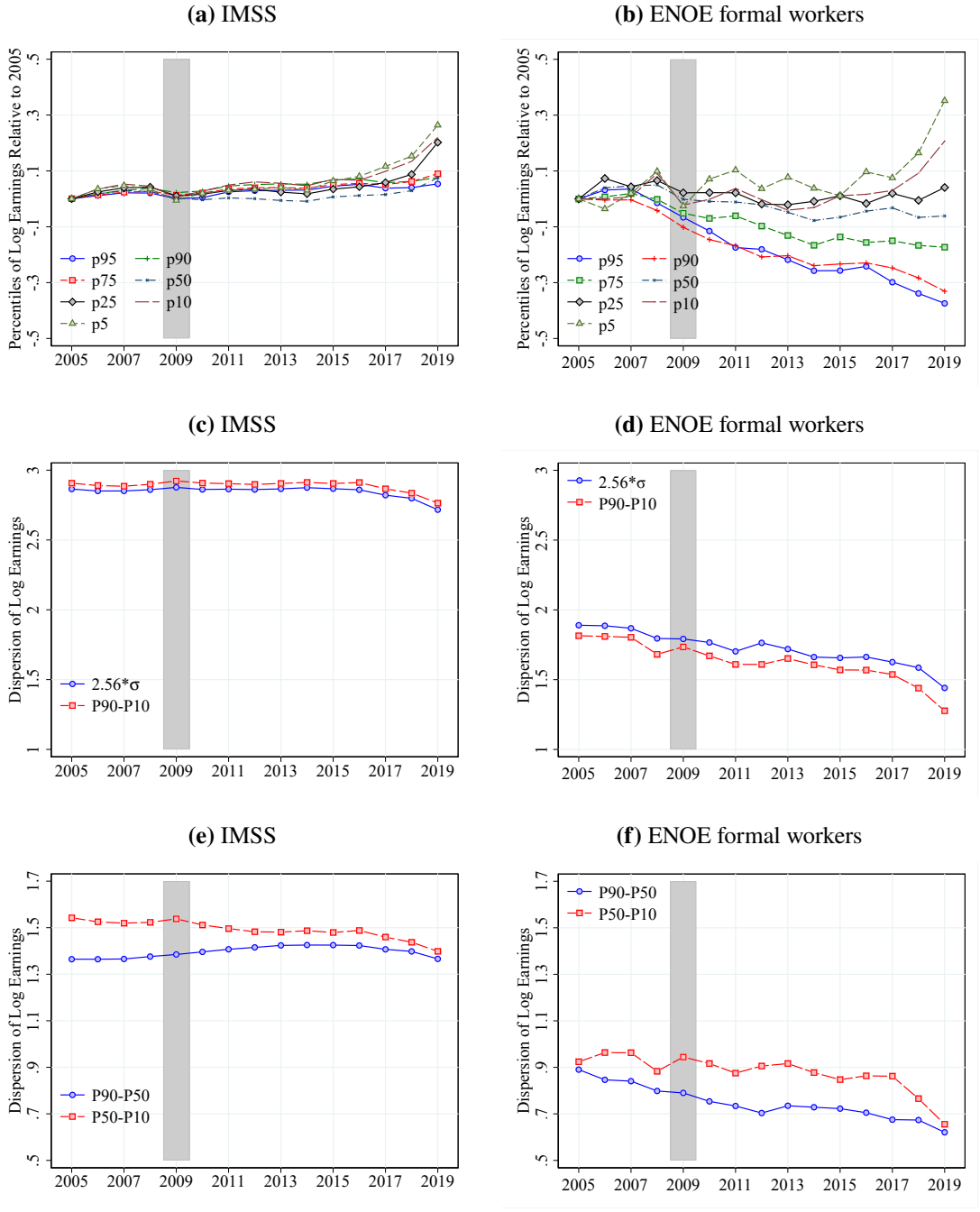
[13] Another well-known household survey run by INEGI that is commonly used to measure poverty and income inequality in Mexico is the Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). We chose to use the

While the underlying structure of the ENOE does not allow for selecting a sample that is identical to the one used in section 3 (i.e. IMSS is a census with a true panel structure while ENOE is a survey sample with a rotating panel structure), we select a sample of workers present in ENOE that is as close as possible to the master sample constructed with the administrative data. We then proceed to calculate statistics analogous to those of sections 3.1 and 3.2. In particular, we only consider individuals who are sampled for 5 consecutive quarters, report a positive income for at least one of the 5 quarters, and report to be employed in an IMSS-affiliated job for at least one of the 5 quarters. This sample of workers is representative of the "formal base population" (i.e. the population of formal workers that is our reference) that more closely resembles that of IMSS. We then apply to this sample a set of restrictions that make it as similar as possible to the IMSS master sample. That is, we impose the same bottom and top coding as in the administrative data, we only consider monetary labor income from the main occupation as this is the earnings measure that more closely matches the one reported in the IMSS records, we restrict the sample to workers aged 25 to 55, and we impose the same "meaningful attachment to the labor force" condition that we imposed to the workers in the administrative data (annual income equivalent to at least 45 days paid at the minimum wage). Since workers in the ENOE report their typical monthly income in each quarter, we calculate their annual income by averaging the monthly income for all the quarters they are observed and then multiplying it by 12. Finally, we recalibrate the expansion factors (weights) originally assigned to the individuals included in the ENOE sample selected by applying these restrictions using a variable that stratifies labor income in terms of number of minimum wages (strata being: at most one minimum wage; between 1 and 2 minimum wages; between 2 and 3; between 3 and 5; more than 5). This ensures that the share of individuals present in each stratum is equal in the original ENOE sample that represents the formal base population and in the final ENOE sample obtained once the restrictions just mentioned are applied.

ENOE instead of the ENIGH for this comparison exercise for several reasons. First, while the ENIGH provides information on various sources of household income, we only need information on (monetary) labor earnings as this is the income variable that is relevant for the comparison with the IMSS-based statistics from section 3. On this front, both surveys provide comparable information. But it is not clear that the ENIGH has a clear advantage over the ENOE as the ENIGH is primarily designed to capture overall wealth, not labor income. Second, the ENIGH is only available every two years, while the ENOE is run every quarter. Third, the ENIGH only provides one data point for labor income for each worker in each year, while the ENOE can provide up to five data points. Using more frequently available information has the advantage of minimizing noise and measurement error in the data. Fourth, the ENIGH went through some important methodological changes during the period studied in this paper that limit comparability across the rounds of this survey. Finally, and most importantly given the nature of the Global Income Dynamics Project, it is impossible to capture any dynamics of income with the ENIGH. Despite all these limitations that lead us to prefer the ENOE, we conducted a similar exercise to the one presented in this section and found that the main trends in inequality that we document with IMSS and ENOE are also observed when using ENIGH data for formal workers and for the total workforce (formal and informal).

We start by comparing the features of the distribution of log earnings calculated with information from IMSS (left) and from ENOE —formal workers only— (right). The upper panel of figure 13 shows the evolution of the percentiles of this distribution. The lowest percentiles, up to approximately P25, have remarkably similar patterns in both IMSS and ENOE, particularly to the extent that they display a significant increase, from 10 to 30% in comparison to 2005, from 2014 onward. The upper percentiles, however, differ starkly when comparing IMSS with ENOE. The ENOE-based top percentiles display very large drops ranging between 15% for P75 and almost 40% for P95 relative to 2005. These changes observed in the ENOE data are rather large and suggest that the household survey may not be an accurate source of information for capturing earnings at the top of the distribution.

Figure 13: Comparison between administrative and survey data: distribution of log real earnings



Note: Based on authors' calculations with data from IMSS and ENOE. Using the sample from the IMSS data and a sample from ENOE (only formal workers with access to social security) constructed to match the IMSS sample as closely as possible, this figure plots against time the following statistics of the distribution of log earnings: (a) IMSS: P5, P10, P25, P50, P75, P90, P95; (b) ENOE formal workers: P5, P10, P25, P50, P75, P90, P95; (c) IMSS: P90–P10 and 2.56*σ; (d) ENOE formal workers: P90–P10 and 2.56*σ; (e) IMSS: P90–P50 and P50–P10; (f) ENOE formal workers: P90–P50 and P50–P10. Shaded areas are recessions.

To understand the source of this discrepancy in the upper part of the log earnings distribution, we analyze a problematic issue that is very common in surveys —non-response. In ENOE, non-response regarding income/earnings for employed workers has increased from 9% in 2005 to more than 25% in 2019. Moreover, the non-response is non-random and concentrated among highly educated formal workers. For university-educated workers, non-response almost doubled between 2005 and 2019, reaching 40% in 2019. For workers with primary education or less, non-response also increased but it never exceeded 20 percent. Formal and urban workers are also more likely to not provide information about their income. More details that document these facts are provided in appendix C. The fact that non-response is so pervasive for specific categories of workers who are usually the top earners, that it has been increasing over time, and that observable characteristics of non-responders have also somewhat changed throughout the sample period, indicates that this issue is likely to be the reason why the ENOE is inaccurate in characterizing the upper part of the earnings distribution.

Looking at the middle and bottom panels of figure 13, we observe a decrease in overall dispersion, and hence a decrease in inequality, that accelerates from 2016 and is more strongly driven by the declining inequality in the lower part of the earnings distribution. These patterns hold with data from both IMSS and ENOE even if the levels are not the same. Our results are reassuring in terms of confirming that administrative and survey data are able to provide a similar picture of the evolution of earnings in the lower percentiles of the distribution. But they also point to the fact that survey data can be particularly unreliable for studying the evolution of the top percentiles, where the non-response tends to be more concentrated and severe, and for computing measures of inequality.

To provide a more comprehensive picture and to better understand how our analysis from section 3 that includes only workers employed in the formal private sector can be representative of all workers in Mexico, we present figures similar to figure 13 in appendix C. These figures analyze the evolution of the percentiles of log earnings and measures of inequality for informal workers and for the whole pool of workers (formal and informal). We find comparable trends indicating that the statistics we calculate for formal workers with both administrative and household survey data are relevant and informative for all employed workers.

Next, we compare the distribution of transitory earnings shocks. Note that, given the short, rotating panel structure of the ENOE (only 5 quarters), we can only construct a measure of one-year income changes with the ENOE, and this measure is inevitably very noisy because we can only use 2 data points for each worker (i.e. information on income in his/her first and last

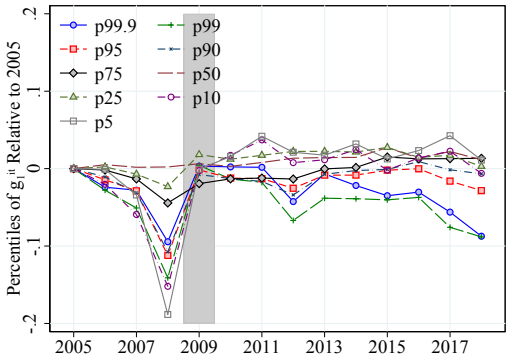
quarter).[14] Figures 14 and 15 show the comparison between the evolution of the percentiles of the distribution of transitory earnings shocks and measures of dispersion, symmetry, and concentration in this distribution, respectively, using data from IMSS and ENOE. We see that the overall trends in the percentiles and moments of the distribution are similar throughout the whole period of analysis with both IMSS and ENOE data. The most notable exception is the financial crisis when some of the deviations from the observed patterns are more evident in the administrative data than in the survey ones. These results suggest that transitory earnings shocks from ENOE and IMSS display more similar patterns than log earnings percentiles. We explore a possible explanation for this finding by verifying how levels of earnings are correlated with one-year earnings changes in the ENOE sample. We calculate that in a given year t their correlation is quite small and statistically significant, ranging between -0.3 and -0.4. But if we focus on the upper part of the distribution of log earnings (from P50 onward), where we know that the non-response is more severe, this correlation is significantly smaller, dropping by more than 30%.[15] This implies that the income level of an individual is not strongly correlated with his/her transitory earnings shocks. That is, individuals with both low and high earnings can experience both small and large earnings changes, and this is particularly true for high earners. Hence, the issues that the non-response causes in the upper part of the distribution of log earnings are not automatically transmitted to the distribution of earnings changes, limiting the extent to which non-response affects the comparability between administrative-based and survey-based measures of temporary earnings shocks.

[14] We use the information on income in the first quarter each worker appears in the survey to construct his/her annual income in year t and the information in his/her fifth (last) quarter to construct his/her annual income in year $t+1$. As we can only use one monthly data point for t and $t+1$ the annualized income, and hence the changes in this income, are very noisy.

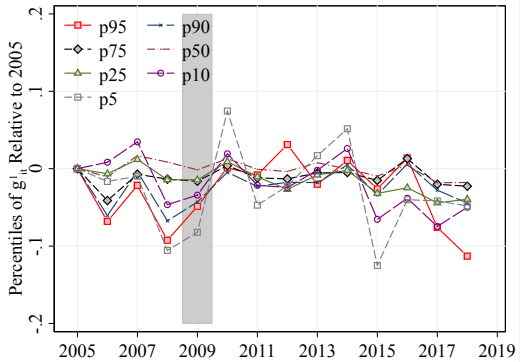
[15] This is consistent with figure 10 that shows that in the administrative data the dispersion of one-year changes in residualized earnings is high in the lowest percentiles of the permanent income distribution but mostly flat everywhere else.

Figure 14: Comparison between administrative and survey data: evolution of the percentiles of the distribution of one-year log earnings changes

(a) IMSS

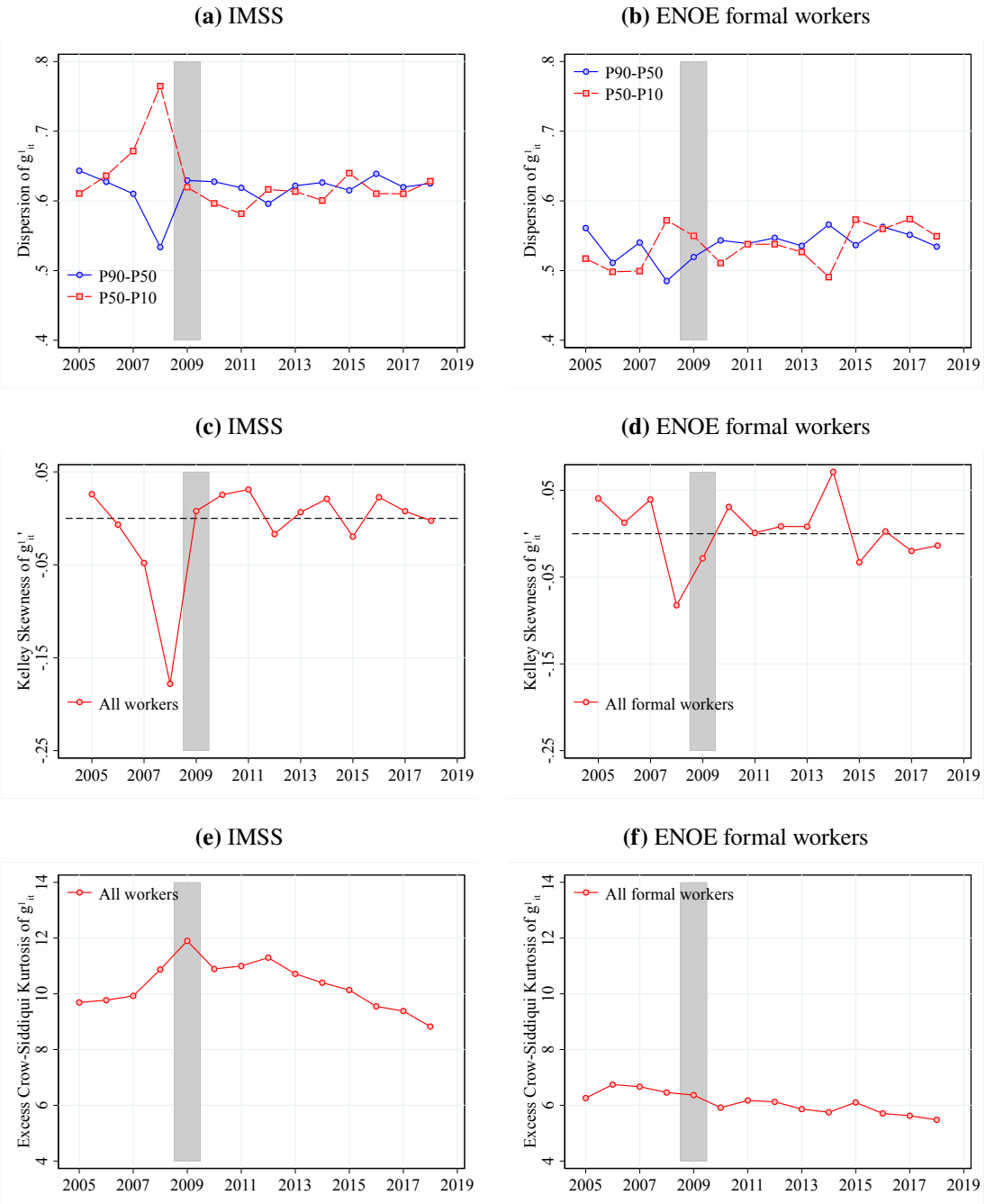


(b) ENOE formal workers



Note: Based on authors' calculations with data from IMSS and ENOE. Using the sample from the IMSS data and a sample from ENOE (only formal workers with access to social security) constructed to match the IMSS sample as closely as possible, this figure plots against time the following statistics of the distribution of one-year log earnings changes: (a) IMSS: P5, P10, P25, P50, P75, P90, P95, P99, P99.9; (b) ENOE formal workers: P5, P10, P25, P50, P75, P90, P95. All percentiles are normalized to 0 in 2005, the first available year. P99 and P99.9 are omitted in panel (b) because, due to the lack of a sufficient number of observations, they are too noisy to be informative. Shaded areas are recessions.

Figure 15: COMPARISON BETWEEN ADMINISTRATIVE AND SURVEY DATA: MEASURES OF DISPERSION, SYMMETRY, AND PEAKDNESS OF THE DISTRIBUTION OF ONE-YEAR LOG EARNINGS CHANGES



Note: Based on authors' calculations with data from IMSS and ENOE. Using the sample from the IMSS data and a sample from ENOE (only formal workers with access to social security) constructed to match the IMSS sample as closely as possible, this figure plots against time the following statistics of the distribution of one-year log earnings changes: (a) IMSS: P90–P50 and P50–P10; (b) ENOE formal workers: P90–P10 and P50–P10; (c) IMSS: Kelly skewness; (b) ENOE formal workers: Kelly skewness; (c) IMSS: Excess Crow-Siddiqui kurtosis; (d) ENOE formal workers: Excess Crow-Siddiqui kurtosis. Shaded areas are recessions.

4.2 Transitions In and Out of Formal Employment and their Impact on Wage Dynamics

A distinctive feature of the formal labor market in Mexico is that there is continuous entry and exit of workers into and out of IMSS-affiliated jobs. For example, during the sample period covered by our analysis, only 8.8% of workers maintained an IMSS-affiliated (formal) job throughout. Additionally, about one fourth of workers have two or more spells of formal employment where a spell is defined as a sequence of contiguous years in which we observe a worker's income within our data (see appendix A for descriptive statistics regarding active employment spells). Thus, in general, most Mexican workers do not seem to have a very strong attachment to formal employment. This tenuous bond can potentially have an important impact on the lifetime earnings and welfare of workers given that informal employment typically implies lower average wages and does not entail social security benefits, nor grants any of the employment protection associated with formal employment.

The movements in and out of the formal employment can be associated with transitions between IMSS-affiliated employment, non-IMSS-affiliated employment (either in the formal public sector or in the informal sector), unemployment, or exit from the labor force. It is important to emphasize that, once workers leave the IMSS database, we are unable to track them and thus cannot ascertain the state into which they transition (nor the state from which they transition back into formal employment). Given that we are focusing on prime-age workers, it is most likely that the bulk of these transitions occur between (private) formal and informal employment, as the informal labor market is one of the most important mechanisms for Mexican workers to smooth income shocks. While information on transitions is not available in the administrative data, the household survey ENOE permits to follow workers across different states (i.e. employed in the formal sector through IMSS; employed in the formal sector through a different social security institution; employed in the informal sector, unemployed; out of the labor force). We use the ENOE to characterize transitions across these states using a sample of workers comparable (in terms of age) to the sample from the IMSS data that we use in the first part of the paper and during the same period (2005–2019), and document that, indeed, the majority of workers transitioning out of IMSS-affiliated jobs move to informality. In particular, we find that in a given quarter, on average, 18% of the male and 19% of female workers leave IMSS-affiliated employment. Out of these transitioning workers, 63% of men and 45% of women end up being informally employed in the next quarter. Similarly, 62% of all male and 46% of all female workers who get employed in an IMSS-affiliated job in one quarter worked in the informal sector in the previous quarter. These

patterns confirm that informality is the most recurrent employment option outside the formal sector for both men and women.[16] As this entry and exit can affect the dynamics of income, we now turn the focus of our analysis to verifying whether spells out of formal employment imply a penalty on earnings upon re-entry and, if so, how large this penalty is and how long it takes for workers to recoup their previous wages.

We restrict our sample to the group of workers with two spells of formal employment.[17] Almost one fifth of all workers who have been formally employed during the sample period have only two spells of formal employment (see appendix A), which implies that they have exited and re-entered after a break that lasted at least one year and hence have only one gap in formal employment. On average, these workers are present in the database 3.2 years before leaving, stay out of formal employment 2.6 years, and then come back for another 3.6 years. The distinguishing features of this subsample are: (i) the proportion of males is slightly larger than for the whole sample (61.2% vs 60.4%); (ii) their average age at the beginning of their first spell of formal employment is slightly lower (31.7 years old versus 32.6 years old); and (iii) their average age at the conclusion of their second spell of formal employment is 40.4 years old, substantially lower than the retirement age. To analyze the potential penalty on earnings implied by this mechanism of entry, exit and subsequent re-entry into formal employment, we compare pre-exit wage trajectories with the post re-entry trajectories. In particular, we adopt an event-study approach around the beginning and the end of the gap in formal employment. In our baseline specification we examine a 3-year event window before and after each worker exits the database. We balance the panel by keeping only workers that can be observed for three consecutive years before leaving the dataset and for three consecutive years after re-entry. The 3-year event window is chosen as the benchmark as it maximizes sample size while mimicking the average duration of these workers' active spells.[18]

The comparison is performed by estimating the following specification, separately by

[16]The survey data also confirm the well-known fact that in Mexico men have a much stronger attachment to the labor force than women. In fact, 79% of male workers who exit IMSS-affiliated jobs remain in the labor force and find employment mainly in the informal sector or in formal jobs not affiliated with IMSS. On the other hand, the share of female workers who remain employed upon leaving IMSS-affiliated jobs is only 60%, with almost three quarters of these workers finding employment in the informal sector, but with a considerable share, about 30%, leaving the labor force (the remaining 10% transitions to unemployment).

[17]The descriptive statistics and estimates relative to this part of the analysis are based on a random sample of 4 million workers obtained from the universe of workers that were present at least one year in the master sample used to carry out the analysis presented in section 3.

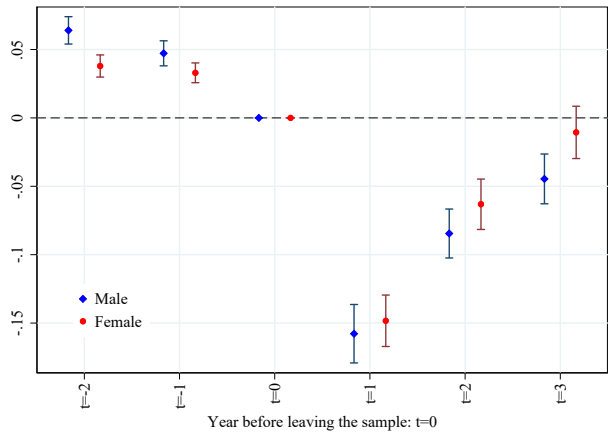
[18]As we widen the analysis window, we lose observations since it is difficult to find many workers that are present in the data many years before leaving as well as after re-entering. As a robustness check, we also analyze a 5-year event window to assess whether wage trajectories change as the event window widens. Inevitably, the 3-year event panel is comprised of a different set of workers than those included in the 5-year event window. Results hold when using the specification with the wider event window (see appendix D).

gender:

$$\ln(w_{it}) = \beta_0 + \sum_{\tau=-2}^{3} \beta_\tau \mathbb{I}_\tau + \sum_{\kappa=1}^{9} \beta_\kappa \mathbb{I}_\kappa + \sum_{\tau=-2}^{3} \sum_{\kappa=1}^{9} \beta_\tau^\kappa \mathbb{I}_\tau \mathbb{I}_\kappa + \gamma_g \mathbb{I}_g + \alpha_e + \alpha_s + \alpha_t + \varepsilon_{it} \qquad (4.1)$$

Here $\ln(w_{it})$ is the logarithm of the average monthly wage of worker i in year t. $\mathbb{I}_\tau = \mathbb{1}[event = \tau]$ are dummy variables for the number of years before or after leaving the sample: $\tau = 0$ is the year right before leaving the sample (i.e. the last year of a worker's first job spell in the formal sector), $\tau = -1$ refers to two years before leaving the sample, $\tau = 1$ is the first year upon re-entering (i.e. the first year of a worker's second job spell in the formal sector), and so on. $\mathbb{I}_\kappa = \mathbb{1}[duration = \kappa]$ is a set of dummy variables that take the value of 1 depending on the k number of years during which the worker was out of formal employment. We consider durations out of formal employment that go from 1 up to 9 years because these are the durations observed in the data. $\mathbb{I}_g = \mathbb{1}[age_{it} = g]$ are dummy variables for the age group to which worker i belongs in year t. α_e, α_s, and α_t are fixed effects for sector of economic activity e and federal state s where the worker is employed, and year t, respectively, and ε_{it} is the error term. Standard errors are clustered at the worker and sector-year levels.

Figure 16: Estimates of wages trajectories (log differences) of workers who exit and re-enter formal employment



Note: Based on authors' estimates with data from IMSS. The figure plots differences of log wages obtained by estimating equation (4.1) using a subsample of workers with only two spells of formal employment. Markers for men and women are positioned to the left and right, respectively, of each event year t. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates.

The coefficients of interest are the β_τ, that capture how wages in different event periods τ compare to the base year, $\tau = 0$, the year right before leaving the sample. A graphical

representation of the estimates of these coefficients is shown in figure 16.[19] On average, wages decrease the year before leaving the sample, both for male and female workers. Men appear to be experiencing a more marked downward trend two years before leaving, while women seem to have more stable wages in the years before exit. Upon re-entry, our results indicate that, in the first year of their second formal job spell, workers suffer a wage penalty of around 15% compared with the wage they earned in the year right before concluding their first spell of formal employment. This holds for men and women, although men seem to fare worse than women in terms of the speed of wage recovery.[20] Wages start to catch up in the second year upon re-entry, but men are still not able to fully regain their pre-exit wages in the third year after re-entering, whereas women are able to recoup almost entirely.

The fact that wages are lower upon re-entry suggests that the wage penalty is likely associated with most of these workers transitioning to lower-paying jobs, possibly in the informal sector, after (temporarily) leaving formal employment. As a consequence, these workers may suffer a negative income shock that the analysis in section 3 is not able to capture. On the other hand, regaining formal employment may also entail a positive income shock. Since our results suggest that the unobserved positive (re-entry) income shock is smaller than the initial negative (exit) one, some of the measures of income dynamics discussed in section 3 may be unable to fully capture the effects of these shocks on earnings as they cannot take into account the impact of transitions in and out of formal employment.

We also characterize the behavior of wages, in levels, during the event window under analysis. We use the estimated coefficients from equation (4.1) and calculate the conditional means of log wages with respect to a worker in the base category of the regression,[21] for each event period τ and each duration of the spell outside of formal employment κ. Specifically, we compute $\mathbb{E}[\ln w_{it}|X = X_0, \tau = t, \kappa = k] = \hat{\beta}_0 + \hat{\beta}_\tau + \hat{\beta}_\kappa + \hat{\beta}^\kappa_\tau$, where X_0 is the base category. Once again standard errors are clustered at the worker and sector-year levels. In addition to the patterns that we already discussed in relation to figure 16, we now see that, on average,

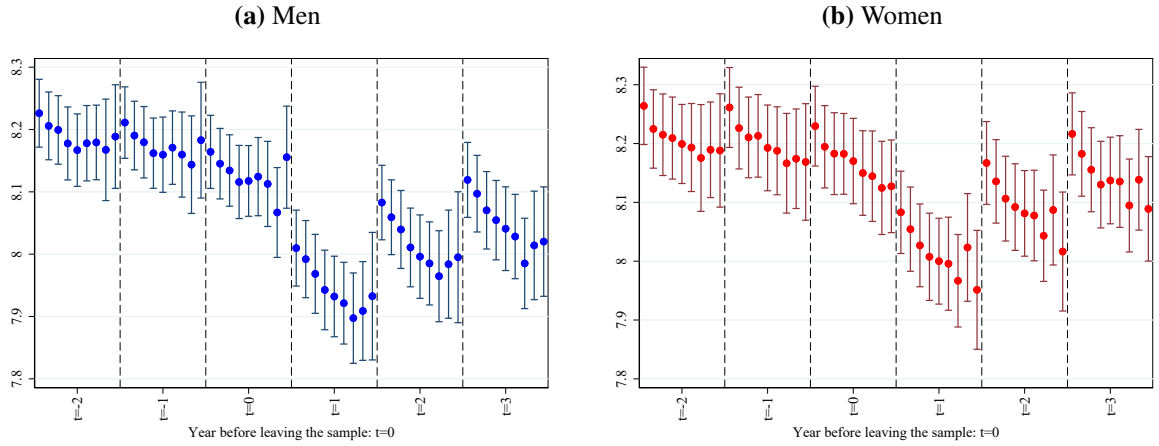
[19] The complete estimation output for this regression is provided in appendix D. In the same appendix we present additional results obtained with a larger sample of workers, those with at least two spells of formal employment, where we consider only their first two spells. We also present results from adding worker fixed effects to (4.1). The main insights are maintained across these different specifications.

[20] We test for differences in the coefficients β^{male}_τ and β^{female}_τ. In the first year upon re-entry the log difference between wages of men and women is not statistically significant, i.e. we cannot reject that $\beta_{1,male} = \beta_{1,female}$. But if we perform a test on the difference in these coefficients along the whole recovery path, namely we jointly test that $\beta_{1,male} = \beta_{1,female}$, and $\beta_{2,male} = \beta_{2,female}$, and $\beta_{3,male} = \beta_{3,female}$, we find that these differences are indeed statistically significant confirming that, overall, men experience a slower wage recovery toward pre-exit levels than women.

[21] The base category of this regression, X_0 is defined as: age=25; sector=agriculture; state=Aguascalientes; year=2005; $\tau = 0$, $\kappa = 1$.

women have higher wages than men upon re-entry[22] and wages for both men and women tend to decrease with the number of years they remained out of the formal sector between their first and second spell of formal employment. Panel (b) of figure 17 shows more clearly that in the third year upon re-entry, women whose stint out of formal employment was relatively brief are able to fully catch up with the wage they earned in the year right before exiting, while men, even those with the shortest duration between spells of formal employment, are still lagging behind.

Figure 17: Estimates of wages trajectories (levels) of workers who exit and re-enter formal employment



Note: Based on authors' estimates with data from IMSS. The figure plots the conditional means of log wages computed as $\mathbb{E}[\ln w_{it}|X = X_0, \tau = t, \kappa = k] = \hat{\beta}_0 + \hat{\beta}_\tau + \hat{\beta}_k + \hat{\beta}^\kappa_\tau$ where X_0 are the observable characteristics of the worker in the base category of regression (4.1) and the tuple (τ, κ) accounts for every possible pair of event τ and duration κ. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates. Standard errors and confidence intervals are obtained with the delta method.

For robustness we verify that our results are not driven by other possibly confounding factors. For instance, we look at the pre-exit and post re-entry wage trajectories of workers whose first spell of formal employment came to an end due to the 2009 financial crisis. We find that the wage patterns we document in our benchmark specification are a general feature of the transitions out and back into formal employment and do not seem to be driven by the specific exit/re–entry that occurred during the financial crisis (see appendix D for a more

[22] We formally test that the difference in the conditional means of female and male workers is different in each event period by running the regression: $\ln(w_{it}) = \beta_0 + \beta_{female}\mathbb{I}_{female} + \sum_{\tau=-2}^{3}\beta_\tau\mathbb{I}_\tau + \sum_{\tau=-2}^{3}\beta_\tau\mathbb{I}_\tau\mathbb{I}_{female} + \alpha_e + \alpha_s + \alpha_t + (\alpha'_e + \alpha'_s + \alpha'_t) \times \mathbb{I}_{female} + \varepsilon_{it}$ and testing whether $\mathbb{E}[\ln(w_{it})|X, \mathbb{I}_{female} = 1, \tau = t] - \mathbb{E}[\ln(w_{it})|X, \mathbb{I}_{female} = 0, \tau = t] = 0$. We confirm that, starting from $\tau = 2$ on, this difference is indeed statistically different than zero.

in-depth discussion of the findings of this exercise).

As an additional robustness check we also run a placebo test by comparing the wage trajectories of workers with two spells of formal employment with the trajectories of the group of workers who stayed in the database the entire period in a fashion very similar to a difference-in-differences event study design. This design is based on assigning placebo exits and duration of spells out of formal employment to individuals who remained formally employed throughout the whole period. Even though we do not intend to interpret these results as causal, they serve as a useful tool for assessing whether the features that we find in the group of workers who leave and re-enter formal employment are peculiar to this group.

To carry out this exercise, we construct a control group comprised of workers who are present in the IMSS data the whole 15 years for which information is available and compare their wage dynamics with those of workers who constitute our treatment group —workers who leave formal employment only once and then come back. We use the Coarsened Exact Matching (CEM) method described in Iacus, King, and Porro [2012] to obtain a balanced sample of the treatment and control groups. Through this methodology we find for 66.3% of workers who could potentially be in the treatment group an exact match in terms of age (age is coarsened into 5-year age groups), gender, sector of economic activity and locality (state) who were observed in their last year before exiting formal employment. Consistent with this methodology, the exact match is chosen randomly among potential candidates, so that we can construct placebo exit events for the individuals in the control group. We assume that each worker in the control group left the database in the same year as his/her match in the treatment group, and that he/she was out of formal employment for the same number of years as his/her treatment group counterpart. Having defined the placebo events for the control group, we then proceed to estimate the following specification:

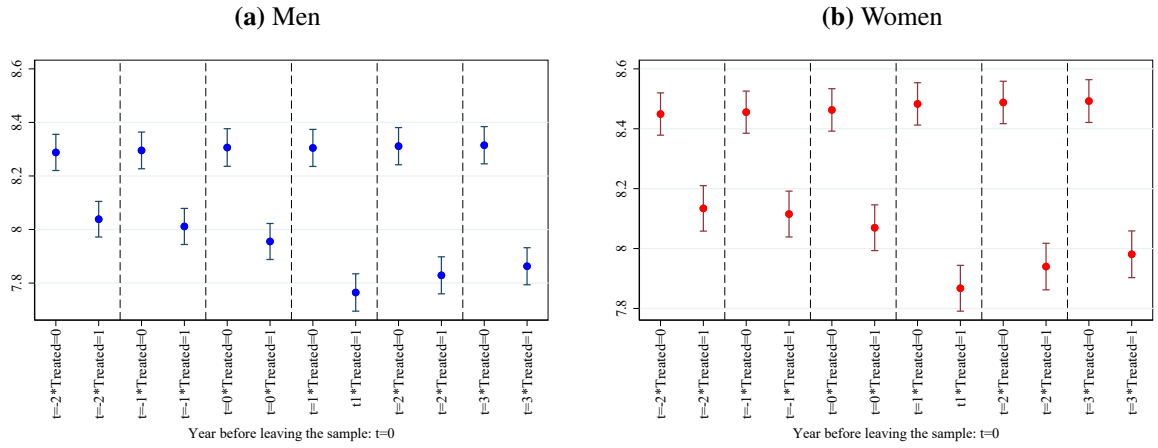
$$\ln(w_{it}) = \sum_{\tau=-2}^{3} \beta_\tau \mathbb{I}_\tau + \delta_{treated} \mathbb{I}_{i,\text{treated}} + \sum_{\tau=-2}^{3} \beta_\tau^{\text{treated}} \mathbb{I}_\tau \mathbb{I}_{i,\text{treated}} + \gamma_g \mathbb{I}_g + \alpha_e + \alpha_s + \alpha_t + \varepsilon_{it} \tag{4.2}$$

where $\mathbb{I}_{i,\text{treated}}$ is a dummy variable that takes the value of 1 if worker i is part of the treatment group and 0 otherwise. We cluster the standard errors at the worker and sector-year levels.

Mean log wages are displayed in figure 18 and are obtained using the estimated coefficients from equation (4.2) to calculate $\mathbb{E}[\ln w_{it} | X = X_0, \tau = t, \mathbb{I}_{i,\text{treated}} = 1] = \hat{\beta}_\tau + \hat{\delta}_{\text{treated}} + \hat{\beta}_\tau^{\text{treated}}$.[23] On average, workers in the control group have higher wages in each event and show a slight upward trend in their wages (more so for women), as opposed to the fall observed in the

[23] The estimation output of equation (4.2) is available in appendix D.

Figure 18: ESTIMATES OF WAGES TRAJECTORIES: TREATMENT VS CONTROL GROUP



Note: Based on authors' estimates with data from IMSS. The figure plots the mean of log wages for the worker in the base category of the regression, conditioning on the worker belonging to the treatment or to the control group, for each event period τ. We compute $\mathbb{E}[\ln w_{it}|X = X_0, \tau = t, \mathbb{I}_{i,\text{treated}} = 1] = \hat{\beta}_\tau + \hat{\delta}_{\text{treated}} + \hat{\beta}_\tau^{\text{treated}}$ using the estimated coefficients from equation (4.2) where we use a control group of workers randomly selected among those who are always present in the IMSS data and a subsample of workers with only two spells of formal employment as the treatment group. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates.

year before exit for workers in the treatment group. These results suggest that the patterns that we estimate in our benchmark specification are not driven by the treatment group merely reflecting a pattern that is also present for workers who do not incur the exit event (i.e. the control group). That is, the wage dynamics for workers who exit and re-enter formal employment can be associated with the fact that these workers spent some time out of the formal sector. For instance, workers who suffer an exit event have lower average pre-exit wages than those who remained continuously attached to a formal job. This suggests that workers at the bottom of the residualized earnings distribution described in section 3.2 may be more likely to exit formal employment. In turn, lower income workers face both a temporary income shock upon exiting formal employment and a more persistent effect on future earnings due to the wage penalty upon re-entry. Thus, entry and exit into and out of formal employment are likely to exacerbate residualized earnings inequality.

4.3 Early Exposure to Informality and Future Earnings

Initial conditions have been often documented to have long-lasting effects on economic outcomes. For example, Chetty, Friedman, and Rockoff [2014] find that being assigned to high

value-added teachers improves students' long-term outcomes in terms of being more likely to attend college and earn higher salaries. Chetty, Hendren, Lin, Majerovitz, and Scuderi [2016] argue that differences in childhood environments play a significant role in shaping gender gaps in adulthood for employment rates and earnings. Oreopoulos, Von Wachter, and Heisz [2012] show that the aggregate labor market conditions faced by workers when first entering the labor market can affect employment and earnings outcomes in the long-term. Motivated by these findings, we investigate in this section whether workers who start their employment life in an informal job face any long-term negative consequences.

The rotating panel structure of the Mexican household survey does not permit to construct a panel of formal and informal workers with the level of detail for their employment trajectories required to study long-term labor market outcomes as in, for example, Schwandt and Von Wachter [2019]. We attempt to partially address this limitation exploiting a special module of the ENOE, the Labor Trajectory Module (MOTRAL, for its acronym in Spanish). The MOTRAL was specifically designed to collect information on labor trajectories of Mexican workers. It was conducted on a subsample from ENOE in the second quarter of both 2012 and 2015. Each round of the MOTRAL has the objective of reconstructing employment trajectories over the previous 5 years (i.e. the 2012 round covers the period 2007–2012 and the 2015 round covers the period 2010–2015). As with the standard household survey, these trajectories are entirely self-reported. In this last part of our analysis we use the MOTRAL to address the following question: how does the formal versus informal status of a worker's first job affect his/her future wages?[24]

A relevant feature of the MOTRAL is that information is reported at the level of an individual's job. That is, the respondent is asked to provide information regarding each of the jobs held during the five-year window covered by the survey. For each job reported by an individual in the MOTRAL, we can observe the month and year in which the individual began that job, his/her position within the job (i.e. employer, employee, own-account worker), the monthly wage, the sector of economic activity associated with the job, the social security institution the individual was affiliated with (if any),[25] and the month and year in which the job ended. In addition, socio-demographic variables can be retrieved since each individual in the MOTRAL is present and can be matched in the corresponding round of the ENOE. Importantly, the information collected for each individual's job is independent of the employment duration in that job so that the within-job dynamics are not reported. This implies, for example, that

[24]While the MOTRAL has its own limitations, the survey does ask respondents about specific details of their first job, even if it took place before the years covered by the survey.

[25]Access to social security benefits is used to determine whether a job can be categorized as formal or informal.

wage dynamics within a job are not observed, we only observe one self-reported income for each individual's job, regardless of the specific spell of employment. This, in turn, requires making some assumptions to construct a panel from the MOTRAL.

We use both rounds of the MOTRAL to construct a panel and define the following key variables at the monthly frequency for the period from January 2007 to June 2015. (i) *Working status*: we assume that an individual is working in a specific month if he/she has at least one job spell in that month; (ii) *Monthly wage*: calculated as total earnings across all active job spells in a given month; and (iii) *Sector of economic of activity*: we assign a sector of economic activity to an individual based on the sector associated with the job in which he/she earned the most monthly income. Since some of these variables in the monthly panel display little month-to-month variation, we aggregate the monthly panel to a yearly frequency and redefine the following variables. (i) *Working status*: dichotomous variable equal to 1 if the individual worked for at least one month within a given year; (ii) *Income*: calculated as the average monthly wage for those months in which the individual was an active worker; and (iii) *Sector of economic activity*: we set a worker's sector of economic activity as the sector of the job in which he/she earned the most income during the year. We also include the year of entry into the labor market, the status (formal or informal) of the first job as well as socio-demographic variables.

Using the annual panel just described, we estimate the following specification, separately for each demographic group $d \in \{M^l, F^l, M^h, F^h\}$:

$$\begin{aligned} \ln(y_{it}) \quad &= \beta_0 + \beta_1 age_{it} + \beta_2 age_{it}^2 + \beta_3 formal_{i0} + \beta_4 (formal_{i0} \times experience_{it}) \\ &+ \beta_5 recession_{i0} + \gamma_t + \gamma_s + \lambda_{it} + \varepsilon_{it} \end{aligned} \tag{4.3}$$

where y_{it} is the average monthly income, age_{it} is the age of individual i at time t, $experience_{it}$ are years of potential experience of individual i at time t, $formal_{i0}$ is a dummy variable that equals one if individual i's first job was formal and zero otherwise, and $recession_{i0}$ is a dummy variable equal to one if individual i's entered the labor force for the first time in a local labor market facing adverse economic conditions. We specifically control for the state of the overall economy in local labor markets to avoid the formal first job variable picking up aggregate economic conditions. γ_t and γ_s are year and sector of economic activity fixed-effects, λ_{it} is a correction term that addresses estimation bias due to possible non-random selection into the sample, and $\varepsilon_{i,t}$ is the error term. The coefficients of interest are β_3 and β_4, that capture the effect that the instance of a worker's first job being formal has on future earnings, both

directly and through the interaction with years of potential experience.[26]

The formal status of a worker's first job may be correlated to other unobserved individual characteristics that could also affect labor earnings, such as unobserved ability. Thus, an OLS regression may not recover the true parameters β_3 and β_4 because of omitted variable bias. We correct for this bias by instrumenting $formal_{i0}$ using measures of local labor market informality rates for young workers for the year in which they were first employed. The identifying assumption is that the informality rate is uncorrelated with these unobserved characteristics of the individual entering the labor market, but it negatively affects the likelihood of starting in a formal job. The results of a battery of tests that corroborate the validity of our instruments of choice are reported in appendix E. Since these rates of informality cannot be calculated prior to 1995, we restrict our panel to workers who entered the labor market in or after 1995.[27]

To construct the correction term, λ_{it}, we follow Heckman [1979] and characterize the selection equation as:

$$l_{i,t} = \Phi(\alpha_0 + \alpha_1 experience_{it} + \alpha_2 experience_{it}^2 + outwork_{it} + \psi_t + \psi_r) + \nu_{it} \tag{4.4}$$

where l_{it} is a dummy variable that equals one if individual i is working in year t, ψ_t and ψ_r are fixed effects for year and region of birth, respectively, and ν_{it} is the error term. The variable $outwork_{it}$ is a variable that counts the number of consecutive years individual i spent without working prior to year t. More specifically, this variable is constructed recursively as: $outwork_{it} = [outwork_{it-1} + 1] \times \mathbb{1}[l_{it-1} = 0]$ with $outwork_{it} = 0$ for the first year in which we have information regarding individual i.

[26] Potential years of experience, $experience_{it}$, is simply the number of years that have elapsed between the year of entry into the labor force and year t. The variable $recession_{i0}$ is an indicator that equals one if the annual growth rate of economic activity in the state in which individual i was born is negative in the year this individual became employed for the first time. Notice that, for lack of more accurate information, we implicitly assume that individuals enter the local labor market for the first time in the state in which they were born. Economic activity is measured with the ITAEE (Indicador Trimestral de la Actividad Económica Estatal), a quarterly index that is a timely macroeconomic indicator at the state level and can be thought as a proxi for a state's GDP. The index is first annualized and then its growth rate is calculated as the annual percentage change in the index.

[27] Approximately 43% of our sample enters the labor market in or after 1995. Figure E.2 in appendix E presents the distribution by year of entry in the whole sample. We also restrict the sample to exclude individuals who are employers, own-account workers, and unpaid employees. These observations are excluded due to potential concerns regarding heterogeneity in the quality of jobs for the workers present in the panel. Since the MOTRAL does not provide additional information regarding job quality, other than the wages earned, the formal/informal nature of the job, and the sector of economic activity, we address this concern by restricting our sample. These excluded observations account for around 20% of the total observations in the panel we construct. Also, as the ENOE is only available from 2005, we use the ENE, the antecedent of the ENOE, from 1995 to 2004 to construct measures of informality rates for these years. The variables used to construct the informality rates are compatible and congruent across the ENE and the ENOE.

The estimation results of the probit model used to construct the Heckman sample selection correction term are reported in table 2.[28] The results show that: (i) potential years of experience have a positive marginal effect on an individual's work status that dissipates as experience is accumulated, with this effect being stronger for men than for women; and (ii) the longer an individual has remained in a spell of non-employment the less likely it is that he/she will be employed in the current period, with this effect being stronger for men than for women. The results also suggest that women have a lower average probability of being employed in any given period. Figure 19 aids in the visualization of these patterns.

Table 2: Heckman selection into employment

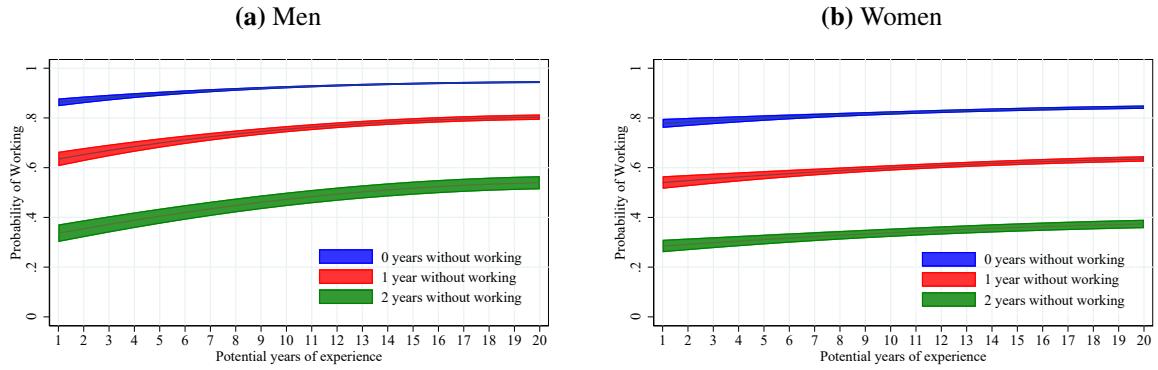
Dependent variable: Labor market participation		
Independent variables:	Men	Women
Experience	0.058^{***}	0.023^{***}
	(0.005)	(0.004)
Experience^2	-0.001^{***}	-0.000^{***}
	(·)	(·)
Out of work	-0.856^{***}	-0.733^{***}
	(0.019)	(0.014)
Constant	0.011	-0.197^{***}
	(0.106)	(0.080)
N. of Observations	20,382	20,058

Note: Based on authors' estimates with data from ENE and ENOE. All specifications include region and year fixed effects. Robust standard errors are reported in parenthesis. Stars indicate significance levels ($^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$).

The estimated coefficients of equation (4.3) are presented in table 3. The main takeaways from this last part of our analysis are the following. Having the first job in the formal sector has a positive level effect on the future wages of high-educated workers, both men and women, which increases with the accumulation of experience. The level effect is of a significant magnitude and particularly so for women: workers that enter the labor market for the first time into a formal job have future wages that, on average, are 45% higher for men and 54% higher for women, relative to the wages of workers whose first job was in the informal sector. Notice also that entering the labor market during a period of contracting economic activity has a long-term negative impact only for high-educated men (about a 5% decrease in future earnings, on average). In contrast, we find that the same effect is positive (10% increase in future earnings, on average) for high-educated women.

[28] Even if the sample for the main regression only considers individuals who entered the labor market in 1995 or after, the probit model in equation (4.4) is estimated with the whole sample.

Figure 19: PROBABILITY OF PARTICIPATING IN THE LABOR MARKET



Note: Based on authors' estimates with data from MOTRAL and ENOE. The figure plots the probability of being employed as a function of accumulated years of experience. Probabilities are calculated assuming average values for fixed effects variables. 95% confidence intervals were estimated using the delta method.

Looking at low educated workers, we find that men can benefit from having a formal first job only through the accumulation of experience. Conversely, having a first job in the formal sector entails a positive level effect on future wages for low-educated women, with no added benefit stemming from the accumulation of potential experience. The effect on average wages for low-educated women is stronger than that estimated for high-educated ones. The fact that potential experience seems to have no relevance in the wage trajectories of low-educated women may be related to potential experience generally overstating women's work experience since it does not account for the time that women spend out of the labor market for child rearing (see De la Cruz Toledo [2014]). Due to strong gender roles in Mexico, women also tend to have more career interruptions than men since they are more heavily burdened with the care of the family. Additionally, our results suggest that entering the labor market during a period of negative growth has no significant effect on future earnings for low-educated workers. This does not imply that these workers do not face depressed wages during the year in which they enter the labor market, but that any potential negative effect on earnings does not persist in the future.

5 Concluding Remarks

Using social security records for millions of Mexican formal sector workers, we have studied the distribution of earnings, mobility patterns in this distribution, and the distribution of earnings changes that characterize the dynamics of earnings. Following a non-parametric

Table 3: Impact of formal first job on future wages

Dependent variable: Log wages						
	Low-educated			*High-educated*		
Independent variables:	Men	Women	All	Men	Women	All
Formal first job	0.022	0.743***	0.185*	0.455***	0.541**	0.257**
	(0.140)	(0.206)	(0.111)	(0.130)	(0.250)	(0.108)
Formal first job×Experience	0.041***	0.011	0.042***	0.037**	0.103***	0.066***
	(0.015)	(0.014)	(0.010)	(0.015)	(0.020)	(0.011)
Age	0.021*	–0.044**	–0.018*	0.027	–0.008	0.027
	(0.012)	(0.022)	(0.011)	(0.023)	(0.031)	(0.017)
Age^2	–0.001***	0.000	–0.000	–0.000	0.000	–0.001**
	(·)	(·)	(·)	(·)	(·)	(·)
Recession	0.041	–0.028	–0.012	–0.051*	0.100*	0.005
	(0.034)	(0.041)	(0.026)	(0.037)	(0.053)	(0.029)
λ	–0.114***	–0.081***	–0.236***	–0.145***	–0.239***	–0.301***
	(0.038)	(0.031)	(0.024)	(0.037)	(0.053)	(0.029)
Constant	8.212***	9.435***	8.881***	7.732***	9.346***	8.488***
	(0.181)	(0.554)	(0.186)	(0.384)	(0.533)	(0.306)
N. of Observations	2,141	1,808	3,949	3,513	3,330	6,843

Note: Based on authors' estimates with data from ENE and ENOE. All specifications include sector of economic activity and year fixed effects. Robust standard errors are reported in parenthesis. Stars indicate significance levels (*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$).

approach, we reach the following conclusions: the distribution of one-year earnings changes displays significant deviations from normality, with these deviations varying over the life-cycle, across the permanent income distribution, and, to a lesser extent, across genders. We find that lower-income and younger workers face, on average, more dispersion in earnings changes, while lower-income and older workers face a distribution of one-year earnings changes with a more pronounced peak. The distribution of earnings changes is not symmetric, but whether it is left or right skewed depends on gender, age, and income: the distribution is most right skewed for lower-income and older women, and it is most left skewed for higher income and younger men. Additionally, the distribution of log earnings displays decreasing dispersion (or inequality) starting in 2015. Finally, we find that upward mobility within the earnings distribution is highest for lower-income and younger workers, while downward mobility is highest for higher-income and younger workers.

After establishing these results, we compare them with results based on a comparable sample of workers from the household survey. The main takeaway of this comparison exercise is that, even if the administrative data do not contain information on an important part of the labor force —informal workers— they are a particularly valuable source of information for studying the top portion of the income distribution and, more importantly, for analyzing income dynamics. Survey data have the advantage of including both formal and informal

workers but suffer from important limitations, such as non-random non-response concentrated among high earners (formal and more educated workers), that may provide an inaccurate picture of important issues such as income inequality and the dynamics of top earnings. We further extend our analysis on the dynamics of earnings of Mexican workers by studying how transitions out of and into formal employment affect earnings and by focusing on the likely role that informality plays in shaping earnings dynamics in Mexico. In this regard, we find that workers who transition out of formal employment are subject to an earnings penalty upon re-entry. This penalty is a cost that workers must bear for three years or more before achieving a level of earnings that is comparable to pre-exit levels. We also document that early exposure to informality, in the form of a worker having his/her first job in the informal sector, proves to have a negative and sizable impact on future earnings.

We hope that our findings can inform future research and policy analysis regarding the Mexican labor market and that studying more in depth its structure and peculiar traits can shed light on the key factors that are germane to the distribution of earnings and earnings shocks in Mexico. Understanding and giving context to these factors is also crucial for performing meaningful cross-country comparisons of these distributions. Future research could benefit from access to tax records that could help overcome the limitations of our analysis for the very top of the earnings distribution. Further work regarding the dual nature of the Mexican labor market, a feature that is also important in other developing countries, and the ways in which earnings dynamics are shaped by workers' transitions across formal and informal employment is also a promising and relevant avenue for future research (see Engbom et al. [2022] for an insightful contribution in this direction).

References

Abowd, J. M., & Stinson, M. H. (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *Review of Economics and Statistics*, *95*(5), 1451–1467.

Acosta, P., Cruces, G., Galiani, S., & Gasparini, L. (2019). Educational upgrading and returns to skills in Latin America: evidence from a supply–demand framework. *Latin American Economic Review*, *28*(1), 1–20.

Arellano, M., Blundell, R., & Bonhomme, S. (2017). Earnings and consumption dynamics: a nonlinear panel data framework. *Econometrica*, *85*(3), 693–734.

Armour, P., Burkhauser, R. V., & Larrimore, J. (2013). Deconstructing income and income inequality measures: a crosswalk from market income to comprehensive income. *American Economic Review*, *103*(3), 173–77.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636.

Binelli, C., & Attanasio, O. (2010). Mexico in the 1990s: the main cross-sectional facts. *Review of Economic Dynamics*, *13*(1), 238–264.

Blanco, A., Díaz de Astarloa, B., Drenik, A., Moser, C., & Trupkin, D. (2022). The evolution of the earnings distribution in a volatile economy: Evidence from Argentina. *Quantitative Economics, forthcoming*.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, *77*(3), 623–685.

Blyde, J., Busso, M., & Romero, D. (2020). *Labor market adjustment to import competition: Long-run evidence from establishment data.* IDB Working Paper Series No. 1100.

Bonhomme, S., & Hospido, L. (2017). The cycle of earnings inequality: evidence from Spanish social security data. *The Economic Journal*, *127*(603), 1244–1278.

Campos Vázquez, R. M. (2013). Efectos de los ingresos no reportados en el nivel y tendencia de la pobreza laboral en México. *Ensayos Revista de Economía (Ensayos Journal of Economics)*, *32*.

Campos Vázquez, R. M., & Lustig, N. (2019). Labour income inequality in Mexico: Puzzles solved and unsolved. *Journal of Economic and Social Measurement*, *44*(4), 203–219.

Campos Vázquez, R. M., & Rodas Milián, J. A. (2020). El efecto faro del salario mínimo en la estructura salarial: evidencias para México. *El Trimestre Económico*, *87*(345), 51–97.

Celik, S., Juhn, C., McCue, K., & Thompson, J. (2012). Recent trends in earnings volatility: Evidence from survey and administrative data. *The BE Journal of Economic Analysis*

& Policy, *12*(2).

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–79.

Chetty, R., Hendren, N., Lin, F., Majerovitz, J., & Scuderi, B. (2016). Childhood environment and gender gaps in adulthood. *American Economic Review*, *106*(5), 282–88.

Chiquiar, D., Covarrubias, E., & Salcedo Cisneros, A. (2017). *Labor market consequences of trade openness and competition in foreign markets.* Banco de México Working Paper Series No. 2017-01.

De la Cruz Toledo, E. (2014). *Women's employment in Mexico.* Doctoral thesis, Graduate School of Arts and Sciences, Columbia University.

De la Cruz Toledo, E. (2018). *Universal preschool and mother's employment.* Working paper.

Engbom, N., Gonzaga, G., Moser, C., & Olivieri, R. (2022). Earnings inequality and dynamics in the presence of informality: The case of Brasil. *Quantitative Economics, forthcoming*.

Engbom, N., & Moser, C. (2021). *Earnings inequality and the minimum wage: Evidence from Brazil.* National Bureau of Economic Research Working Paper No. 28831.

Esquivel, G., Lustig, N., & Scott, J. (2010). A decade of falling inequality in Mexico: market forces or state action? In L. F. Lopez Calva & N. Lustig (Eds.), *Declining inequality in Latin America: A decade of progress?* (pp. 175–217). Brookings Institution and UNDP, Washington D.C.

Goldin, C., & Katz, L. F. (2010). *The race between education and technology*. harvard university press.

Guvenen, F., Kaplan, G., Song, J., & Weidner, J. (2017). *Lifetime incomes in the United States over six decades.* National Bureau of Economic Research Working Paper No. 23371.

Guvenen, F., Karahan, F., Ozkan, S., & Song, J. (2021). What do data on millions of US workers reveal about life-cycle earnings dynamics? *Econometrica, forthcoming*.

Guvenen, F., Ozkan, S., & Song, J. (2014). The nature of countercyclical income risk. *Journal of Political Economy*, *122*(3), 621–660.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 153–161.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 1–24.

Jaume, D. (2021). The labor market effects of an educational expansion. *Journal of*

Development Economics, *149*, 102619.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. *American Economic Review*, *105*(3), 1177–1216.

Kaplan, D. S., & Novaro, F. P. A. (2006). El efecto de los salarios mínimos en los ingresos laborales de México. *El Trimestre Económico*, 139–173.

Katz, L. F., & Murphy, K. M. (1992). Changes in relative wages, 1963–1987: Supply and demand factors. *The quarterly journal of economics*, *107*(1), 35–78.

Kumler, T., Verhoogen, E., & Frías, J. A. (2020). Enlisting employees in improving payroll-tax compliance: Evidence from Mexico. *The Review of Economics and Statistics*, *102*(5), 881–896.

Levy, S. A. (2018). *Under-rewarded efforts: The elusive quest for prosperity in Mexico*. Inter-American Development Bank.

Lustig, N., Lopez-Calva, L. F., & Ortiz-Juarez, E. (2013). Declining inequality in Latin America in the 2000s: The cases of Argentina, Brazil, and Mexico. *World development*, *44*, 129–141.

Mathy, G. P. (2020). How much did uncertainty shocks matter in the Great Depression? *Cliometrica*, *14*(2), 283–323.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199–226.

National Research Council. (2013). *Nonresponse in social science surveys: A research agenda*. National Academies Press.

Oreopoulos, P., Von Wachter, T., & Heisz, A. (2012). The short-and long-term career effects of graduating in a recession. *American Economic Journal: Applied Economics*, *4*(1), 1–29.

Schwandt, H., & Von Wachter, T. (2019). Unlucky cohorts: Estimating the long-term effects of entering the labor market in a recession in large cross-sectional data sets. *Journal of Labor Economics*, *37*(S1), S161–S198.

Stock, J. H., Yogo, M., et al. (2005). Testing for weak instruments in linear IV regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, *80*(4.2), 1.

A Appendix: IMSS Data and Descriptive Statistics from the Master Sample

In this appendix we provide additional information regarding the structure and the specific features of the administrative data used to carry out the descriptive analysis in section 3 and report relevant summary statistics of our master sample that should facilitate the comparison of the results across countries.

Our administrative data, the IMSS data, are available on a monthly basis from January 2005 to December 2019 and cover, approximately, between 13 million workers at the start of the sample period and 20 million workers toward the end. The information available for each worker is:[29]

- ◇ Social Security Number (SSN)
- ◇ Unique population registry* (CURP for its acronym in Spanish)
- ◇ Gender
- ◇ Type of employment (permanent vs temporary contract)[30]
- ◇ Daily (taxable) wage
- ◇ Employer id
- ◇ Firm tax id* (RFC for its acronym in Spanish)
- ◇ Sector of economic activity
- ◇ Geographic location of employer (county where the employer registers the employees with IMSS)

Although not directly provided by IMSS, the worker's SSN provides enough information to infer the year of birth (age) and the year of first enrollment in social security.[31]

While our social security data contain sufficient information to characterize the patterns of income dynamics and inequality for Mexican workers, there are a few issues regarding their limitations that are relevant for understanding how they may compare to or differ from administrative records and/or employer-employee matched data in other countries. First, the IMSS data do not contain any information on workers employed in the informal sector which

[29] These fields of information correspond to those that IMSS has agreed to share with the General Directorate of Economic Research at Banco de México. The fields identified with asterisks are only available from November 2018 onward.

[30] Temporary contracts are those that are specified with start and end dates, while permanent contracts do not include a pre-specified end date.

[31] For some of the observations, the age variable (inferred from the SSN) corresponds to an age that cannot possibly be correct. These observations represent a negligible fraction of monthly observations and are eliminated once the age restrictions are applied for constructing the master sample.

means, as already mentioned, that they miss a very substantial fraction of the Mexican labor force. Second, two additional issues are worthy of mention:

a. Employer id vs Firm id. In the IMSS data, the employer id does not correspond to the firm id as it is usually the case in employer-employee matched datasets. The Mexican social security system allows firms to have multiple "registros patronales" (i.e. employer ids) that are used to register their workers with IMSS. The same firm could use multiple employer ids for several reasons such as operating multiple plants, or employing groups of workers with different risk profiles and there is no official source of information providing a concordance between employer ids and the firms these belong to.[32] The variable firm id in the data corresponds to the id with which each firm is registered in the the "Registro Federal de Contribuyentes" (RFC), a tax identity code assigned by the Servicio de Administración Tributaria, the Mexican tax authority. This code is used by both firms and individuals engaging in economic activities subject to taxes. Analogous to the case of the employer id, firms may legally register multiple RFCs and there is no information regarding which RFCs belong to the same firm.[33]
b. Demographics. Mexican social security data do not provide information on a worker's educational attainment, occupation of employment, or foreign-born status, nor allows for identifying households (neither partners, parents, nor children of a worker). IMSS data cannot be linked with other sources of information, such as household surveys, that contain some of these demographic characteristics.

We now turn to the specifics of the sample construction and sample statistics. Original IMSS records are collapsed to a yearly frequency by summing all the wage observations for a given worker within the year.[34] For the period 2005–2019, this results in over 315 million worker-year pairs with between 17 and 26 million observations per year for workers aged 14–75 years old. Imposing age restrictions on the sample to only include workers aged 25–55, which is the relevant group for all the results of section 3, between 23 and 26% of yearly observations and 24% of the observations in the whole sample are lost. Excluded observations mainly consist of workers aged 24 and below (see figure A.1).

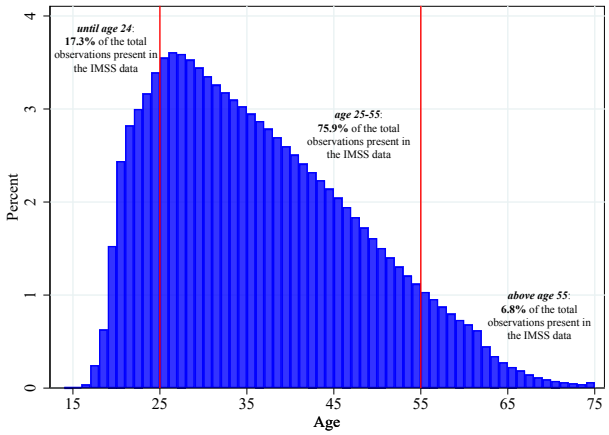
Regarding the composition by gender, the age-restricted sample is consistent with the composition of the original data: on average, throughout the sample period, 63% of obser-

[32] In Mexico, social security contributions are paid based on the risk profile of the worker's occupation.

[33] For example, it could be the case that a firm uses one RFC for its taxable domestic operations and another RFC for its foreign sector operations. But, there is no information on how many RFCs each firm possesses and how it uses them.

[34] At this stage, the only observations that are dropped from the original records are those with a missing value for wage. These observations represent a negligible share of monthly observations.

Figure A.1: AGE DISTRIBUTION IN THE MASTER SAMPLE



Note: Based on authors' calculations with data from IMSS. The distribution is calculated over the entire sample, without age restrictions, consisting of 315 million worker-year pairs.

vations are men, with the share of women rising steadily from 35% in 2005 to 39% in 2019. This implies that the absolute number of men grew by 50% between 2005 and 2019, while the absolute number of women grew by 76% during the same period.[35]

Within the group of workers in question (25–55 years of age), the bulk of the observations are concentrated among workers aged 30 to 44 that, on average, command 55% of yearly observations. By splitting the observations into three age groups, we see that there is a significant change in the distribution of yearly observations across these groups, with a noticeable increase in the participation of the eldest workers (45–55), at the expense of the participation of both the youngest workers (25–29) and workers aged 30 to 44.

Table A.1: GENDER AND AGE COMPOSITION OF THE AGE-RESTRICTED SAMPLE

Year	by Gender		by Age group in % share		
	Men	Women	[25–29]	[30–44]	[45–55]
2005	65.2	34.8	25.1	56.3	18.6
2010	63.6	36.4	23.3	55.8	20.9
2015	63.0	37.0	22.9	54.1	23.0
2019	61.3	38.7	22.6	52.4	25.0

Note: Based on authors' calculations with IMSS data.

An important characteristic of our administrative data is the frequent movement of workers in and out of jobs affiliated with social security. These movements represent transitions

[35] In the original monthly records spanning over January 2005 to December 2019, roughly 64% of observations are men, with the share of women rising steadily throughout the sample, from about 35% at the outset to 37% by the end of the period.

between formal employment and non-formal employment, with this latter state being either employment in the government, employment in the informal sector, unemployment, or exit from the labor force. We highlight some relevant statistics regarding these transitions. Based on an analysis conducted with a random sample of 4 million workers aged 25 to 55 we document that (see also table A.2):

i. 8.8% are present during the entire sample period from 2005 to 2019.
ii. 75.4% have only one active spell that has an average duration of 5.9 years.[36]
iii. 18.9% have only two active spells, the first of these having an average duration of 3.2 years, and the second having an average duration of 3.6 years.[37]
iv. 24.6% of workers have at least one inactive spell during the sample period. Among these, 76.9% have only one inactive spell that has an average duration of 2.9 years.
v. The average duration of the first active spell is 5.2 years, regardless of the total number of active spells.
vi. The average duration of the first inactive spell is 2.7 years, regardless of the total number of inactive spells.

Table A.2: DISTRIBUTION OF WORKERS BY NUMBER OF ACTIVE JOB SPELLS IN THE FORMAL SECTOR

N. of spells	Share %		
	Men	Women	All
1	74.5	76.8	75.4
2	19.2	18.5	18.9
3	5.1	4.0	4.7
4	1.1	0.6	0.9
5 or more	0.2	0.1	0.1

Note: Based on authors' calculations with a random sample from IMSS data consisting of 4 million workers aged 25 to 55.

The share of individuals with only one active spell as formal workers, 75.4%, is the result of a combination of: workers that stayed in the database during the whole sample period (8.8%); workers who entered formal employment after 2005 and continuously kept a formal job until 2019, (29.2%); and workers who entered in or after 2005, ended their formal employment relationship before 2019, and did not regain formal employment before or in 2019, i.e. they did not come back into the database (37.4%). We refer to the first two groups of workers as "stayers" and to the third group as "leavers". The distribution of workers that

[36] Recall that a spell is defined as a sequence of contiguous years in which we observe the worker's income (wage).

[37] 94% of workers have at most two active spells.

have only one formal spell according to these two categories is characterized in table A.3. Stayers and leavers are almost equally split in the sample and, in general, Mexican workers tend to have a tenuous connection with employment in the formal sector, with this being more evident for women.

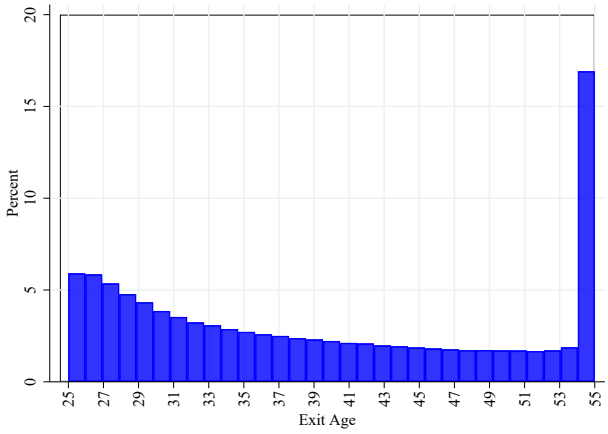
Table A.3: DISTRIBUTION OF WORKERS WITH ONLY ONE JOB SPELL IN THE FORMAL SECTOR

	Share %		
	Men	Women	All
Stayers	51.9	48.2	50.4
Leavers	48.1	51.8	49.6

Note: Based on authors' calculations with a random sample from IMSS data consisting of 4 million workers aged 25 to 55.

Figure A.2 shows the distribution of leavers by age of exit. About 16.9% of these workers leave the IMSS dataset when they are 55 years old. That is, they exit either because they reach the upper age limit we imposed to be included in the master sample or because they retire altogether. At the other extreme, a significant share of workers exit formal employment at a young age: 30.1% of all workers that leave are 30 years old or younger.

Figure A.2: DISTRIBUTION OF LEAVERS BY AGE OF EXIT

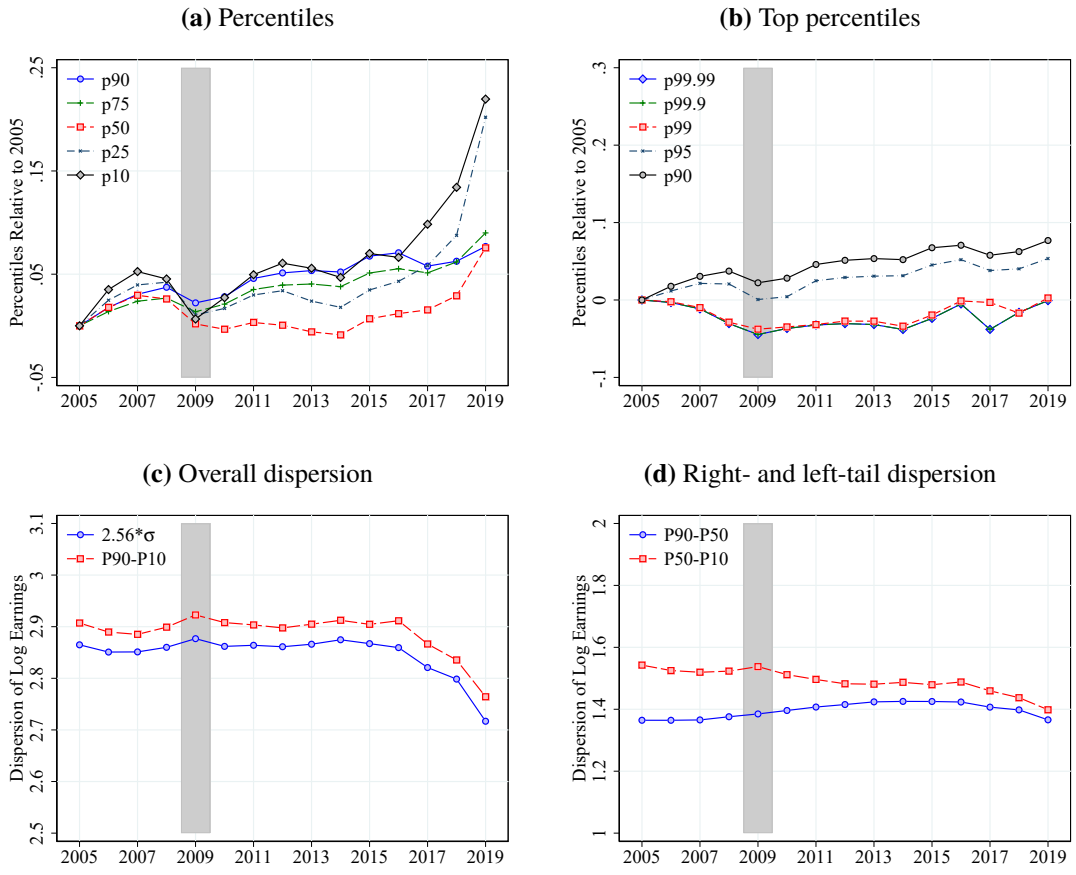


Note: Based on authors' calculations with data from IMSS.

B Appendix: Additional Results for Inequality, Mobility, and Income Dynamics

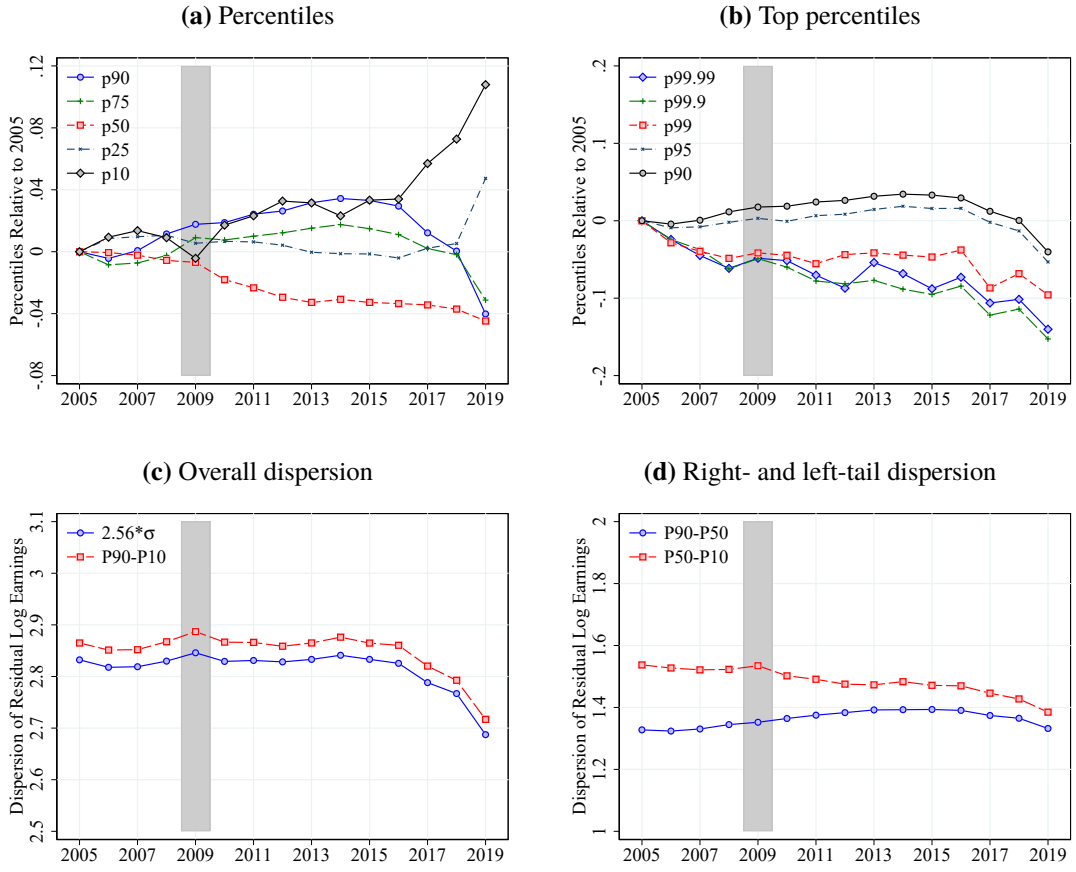
In this appendix we present additional results that complement those presented in section 3 of the main text.

Figure B.1: Distribution of earnings in the population



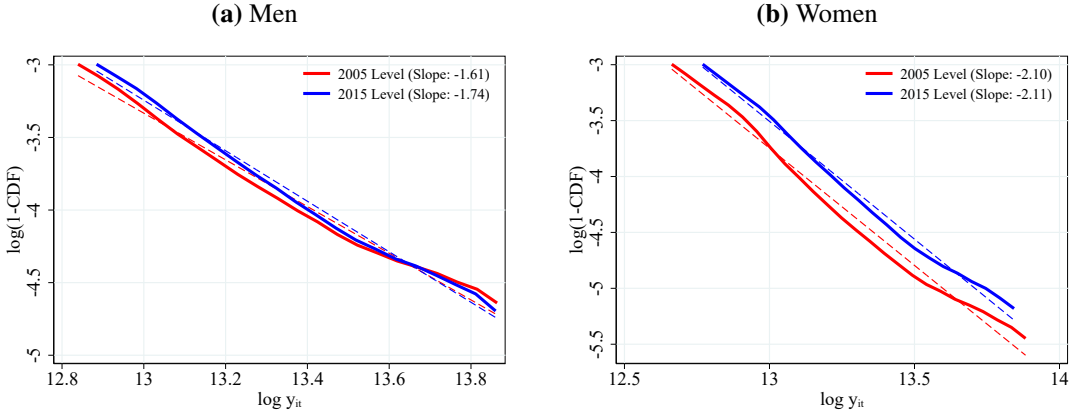
Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time the following statistics of the distribution of log (real) earnings for the whole population: (a) P10, P25, P50, P75, P90; (b) P90, P99, P99.9, P99.99; (c) P90–P10 and $2.56 * \sigma$ that corresponds to the P90–P10 differential for a Gaussian distribution; (d) P90–P50 and P50–P10. Since the data are top coded the percentiles above P95 are imputed by fitting a Pareto distribution around the top code. All percentiles are normalized to 0 in 2005, the first available year. Shaded areas are recessions.

Figure B.2: DISTRIBUTION OF RESIDUAL EARNINGS IN THE POPULATION AFTER CONTROLLING FOR AGE



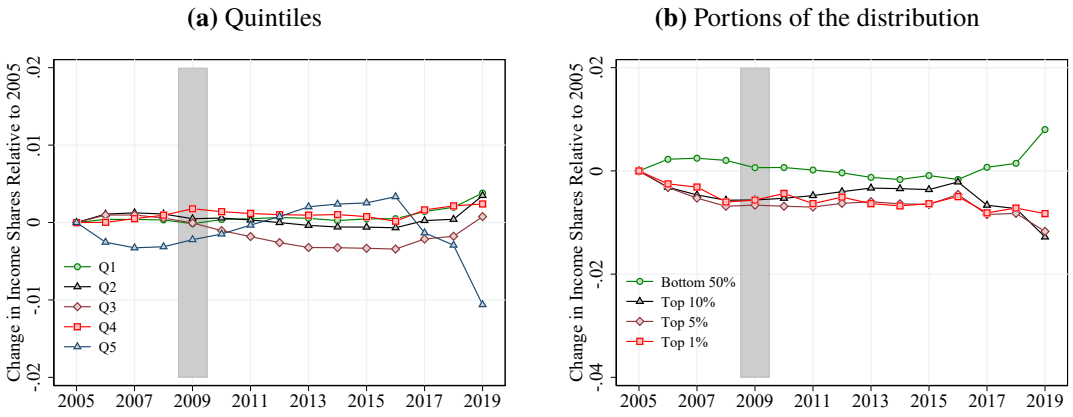
Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time the following statistics of the distribution of residual earnings for the whole population: (a) P10, P25, P50, P75, P90; (b) P90, P99, P99.9, P99.99; (c) P90–P10 and $2.56 * \sigma$ that corresponds to the P90–P10 differential for a Gaussian distribution; (d) P90–P50 and P50–P10. Residual earnings are obtained regressing log earnings against a full set of age dummies, separately by gender and year, and are computed to avoid trends being affected by individuals being at different stages of their life cycles, or by the business cycle. Shaded areas are recessions.

Figure B.3: Top income inequality: Pareto tail at top 5%



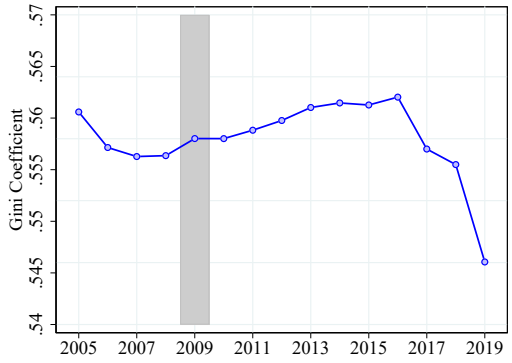
Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against log earnings and for selected years in the sample the following variables: (a) Men: log counter cumulative distribution of earnings; (b) Women: log counter cumulative distribution of earnings. The log counter cumulative distribution is calculated as log(1–CDF). The estimated tail index for a power law distribution in the upper tail, is reported in parentheses. The dotted lines are linear trends. Since the data are top coded and the top percentiles imputed, the figure reports top income inequality at the top 5% of the distribution of log earnings.

Figure B.4: Changes in income share relative to 2005



Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time changes in the distribution of income shares relative to 2005 for the whole population: (a): changes in the quintiles of the income shares distribution; (b) changes in selected portions of the income shares distribution. Quintiles are normalized to 0 in 2005, the first available year. Since the data are top coded and the top percentiles imputed, the figure reports changes in income shares only up until the top 1% of the distribution. Shaded areas are recessions.

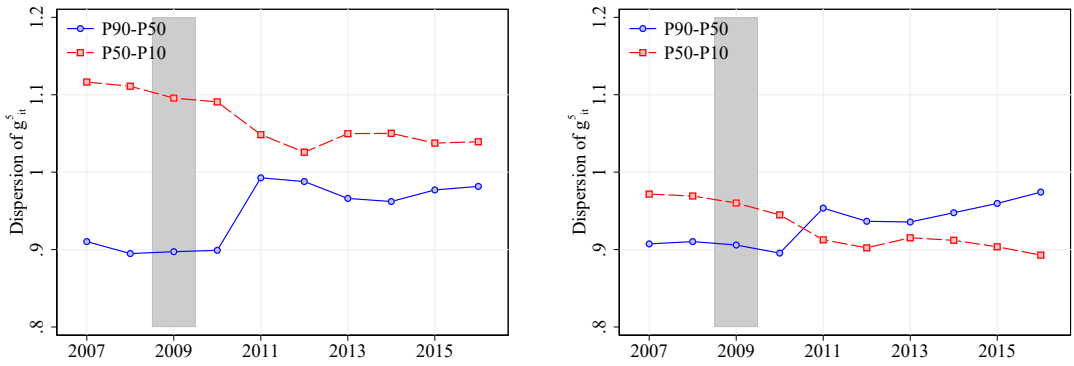
Figure B.5: Evolution of the Gini coefficient



Note: Based on authors' calculations with data from IMSS. Using the cross-sectional dimension of the sample, this figure plots against time the Gini coefficient for the whole population. A Gini coefficient equal to 0 expresses perfect equality in the income distribution, while a Gini coefficient equal to 1 expresses maximal inequality. Shaded areas are recessions.

Figure B.6: Dispersion of five-year earnings changes

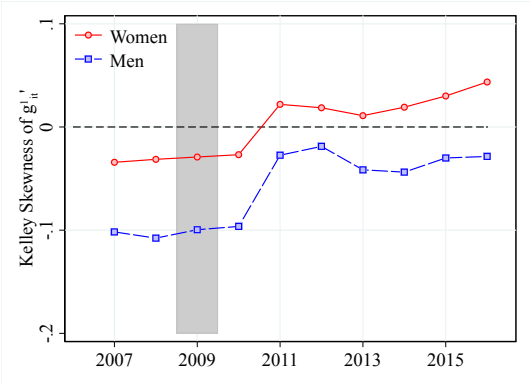
(a) Men **(b)** Women



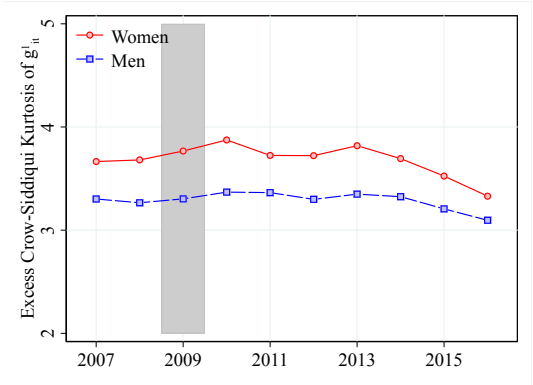
Note: Based on authors' calculations with data from IMSS. Using the time-series dimension of the sample, this figure plots against time the following measures of top- and bottom-tail dispersion of the distribution of five-year earnings changes: (a): Men: P90–P50 and P50–P10 differentials; (b) Women: P90–P50 and P50–P10 differentials. Shaded areas are recessions.

Figure B.7: SKEWNESS AND KURTOSIS OF FIVE-YEAR EARNINGS CHANGES

(a) Kelley skewness

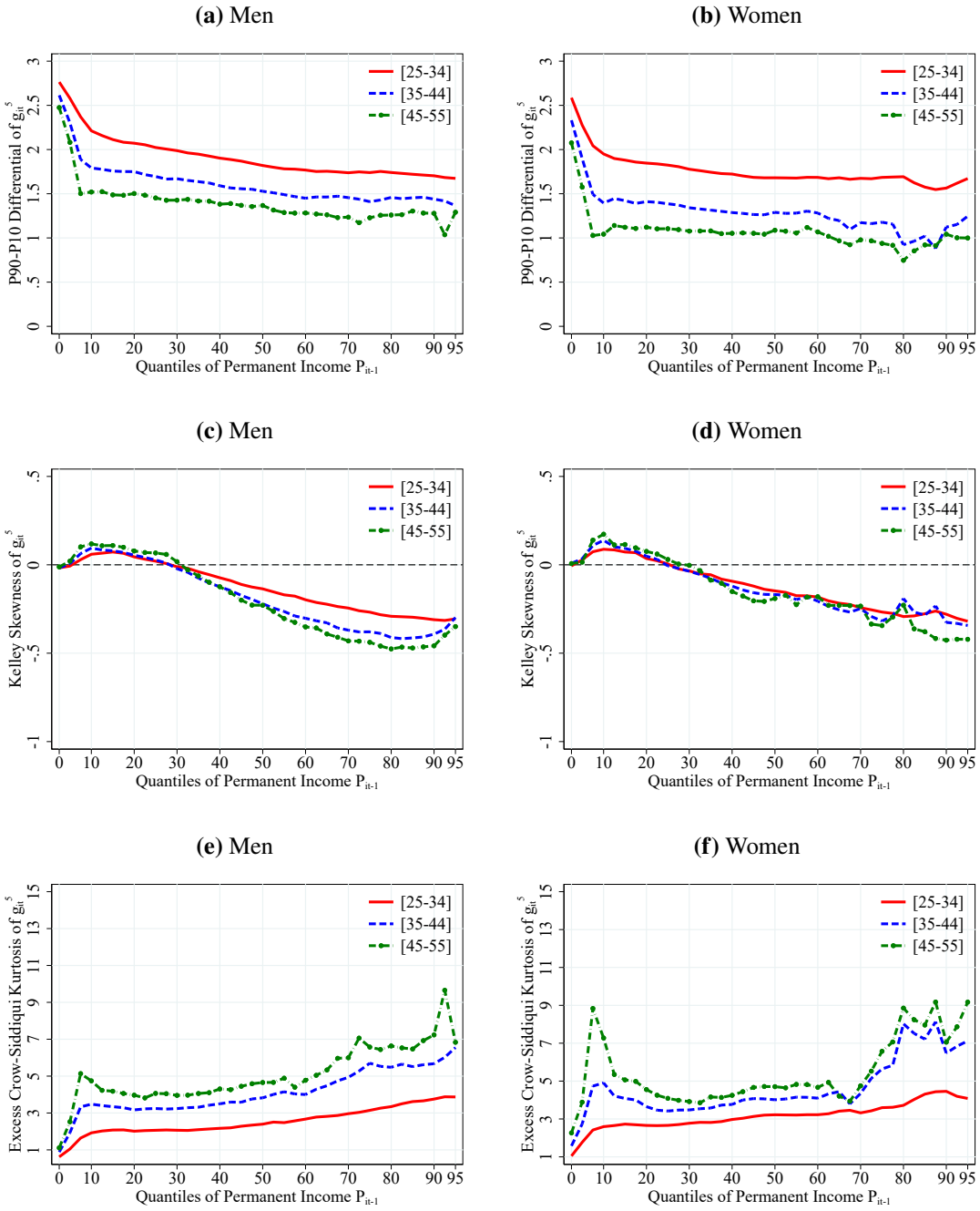


(b) Excess Crow-Siddiqui kurtosis



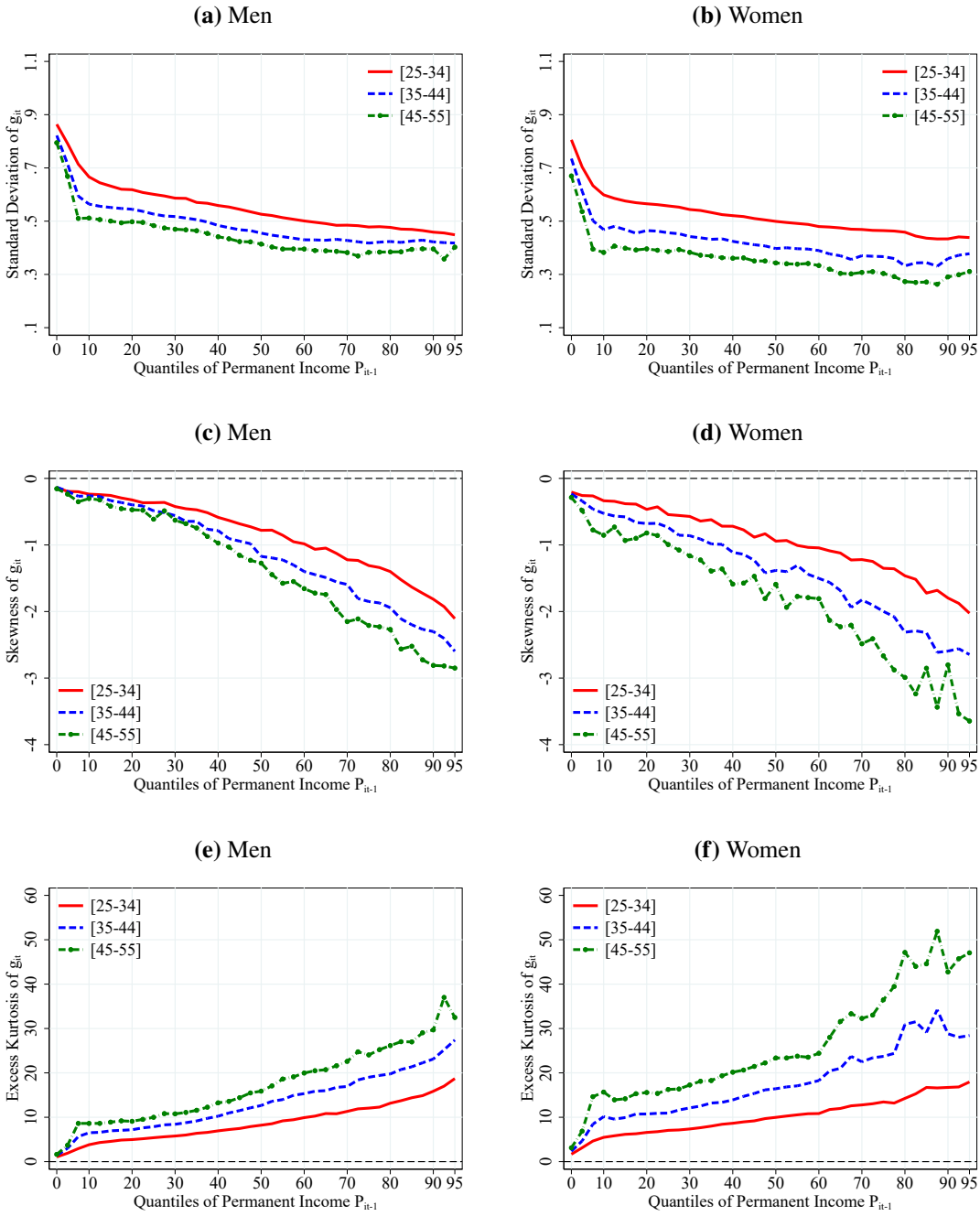
Note: Based on authors' calculations with data from IMSS. Using the time-series dimension of the sample, this figure plots against time the following higher order moments of the distribution of five-year earnings changes: (a) Men and Women: Kelley skewness calculated as $\frac{(P_{90}-P_{50})-(P_{50}-P_{10})}{P_{90}-P_{10}}$; (b) Men and Women: Excess Crow-Siddiqui kurtosis calculated as $\frac{P_{97.5}-P_{2.5}}{P_{75}-P_{25}} - 2.91$, where the first term is the Crow-Siddiqui measure of Kurtosis and 2.91 corresponds to the value of this measure for the Normal distribution. Shaded areas are recessions.

Figure B.8: DISPERSION, SKEWNESS, AND KURTOSIS OF FIVE-YEAR EARNINGS CHANGES



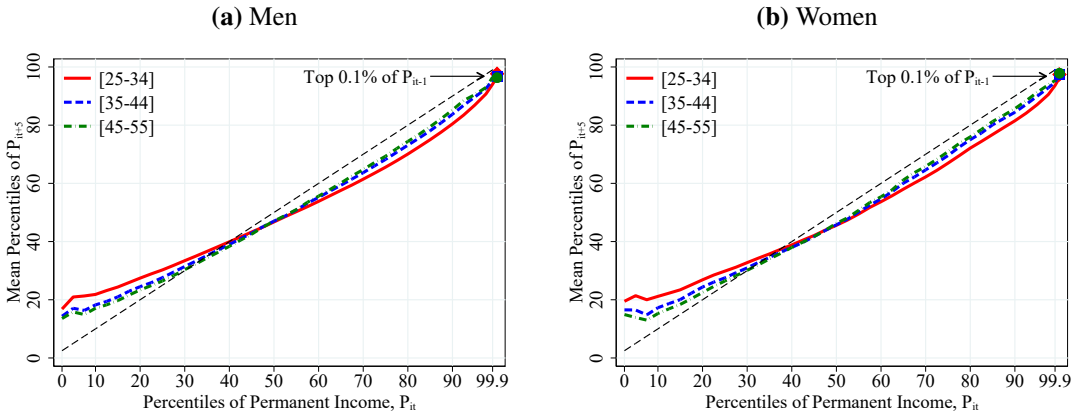
Note: Based on authors' calculations with data from IMSS. Using the worker heterogeneity dimension of the sample, this figure plots against percentiles of the permanent income distribution, and for three different age groups, the following moments of the distribution of five-year earnings changes: (a) and (b) Men and Women: P90–P10 differential; (c) and (d) Men and Women: Kelley Skewness; (e) and (f) Men and Women: Excess Crow-Siddiqui kurtosis. The permanent income is calculated aggregating over a period of 15 years, the maximum number of years available in our sample, from 2005 to 2019. Since the data are top coded the percentiles of the permanent income distribution are plotted only until P95.

Figure B.9: STANDARDIZED MOMENTS OF EARNINGS CHANGES



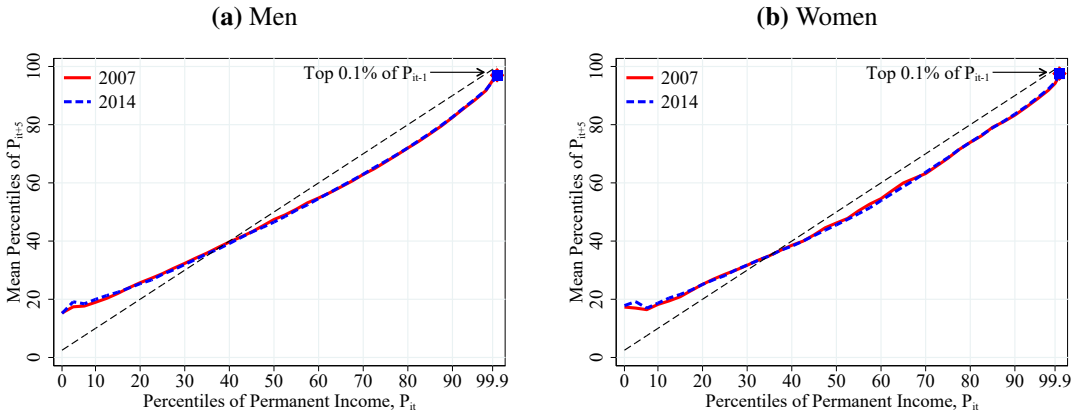
Note: Based on authors' calculations with data from IMSS. Using the worker heterogeneity dimension of the sample, this figure plots against percentiles of the permanent income distribution, and for three different age groups, the following standardized moments of the distribution of one-year earnings changes: (a) and (b) Men and Women: Standard deviation; (c) and (d) Men and Women: Coefficient of skewness; (e) and (f) Men and Women: Excess kurtosis. Excess kurtosis calculated as $\gamma - 3$, where γ is the standard measure of kurtosis (i.e. fourth standardized moment) and 3 corresponds to the value of this measure for the Normal distribution. The permanent income is calculated aggregating over a period of 15 years, the maximum number of years available in our sample, from 2005 to 2019. Since the data are top coded the percentiles of the permanent income distribution are plotted only until P95.

Figure B.10: EVOLUTION OF 5-YEAR MOBILITY OVER THE LIFE CYCLE



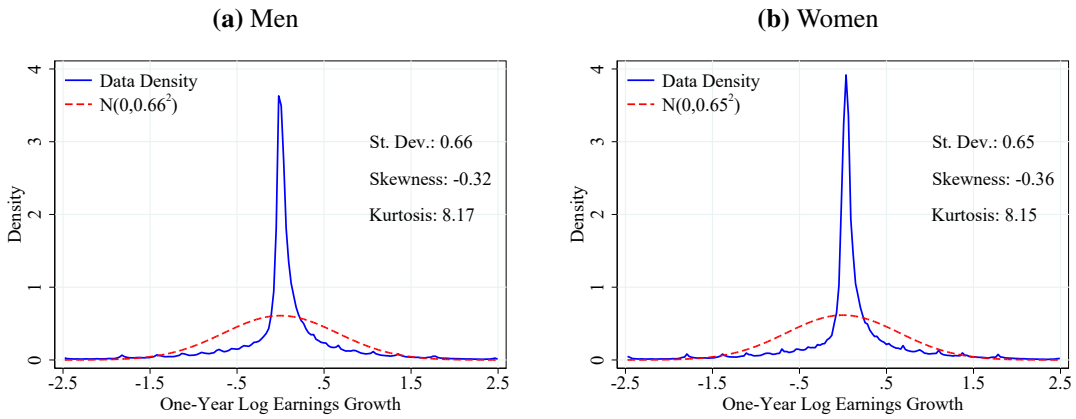
Note: Based on authors' calculations with data from IMSS. The figure shows average rank-rank short-term (5-year) mobility for male (a) and female (b) workers of different ages.

Figure B.11: EVOLUTION OF 5-YEAR MOBILITY OVER TIME



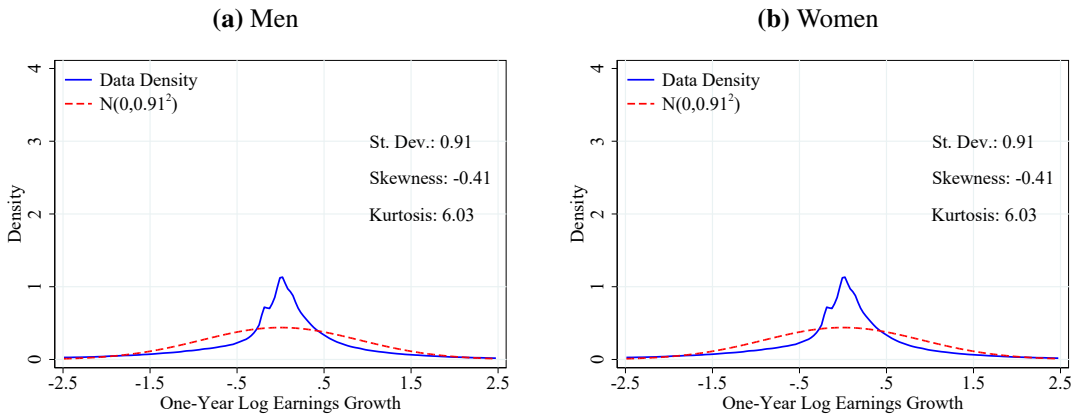
Note: Based on authors' calculations with data from IMSS. The figure shows average rank-rank short-term (5-year) mobility for male (a) and female (b) workers in selected years of the sample, 2007 and 2014.

Figure B.12: EMPIRICAL LOG-DENSITIES OF ONE-YEAR EARNINGS CHANGES



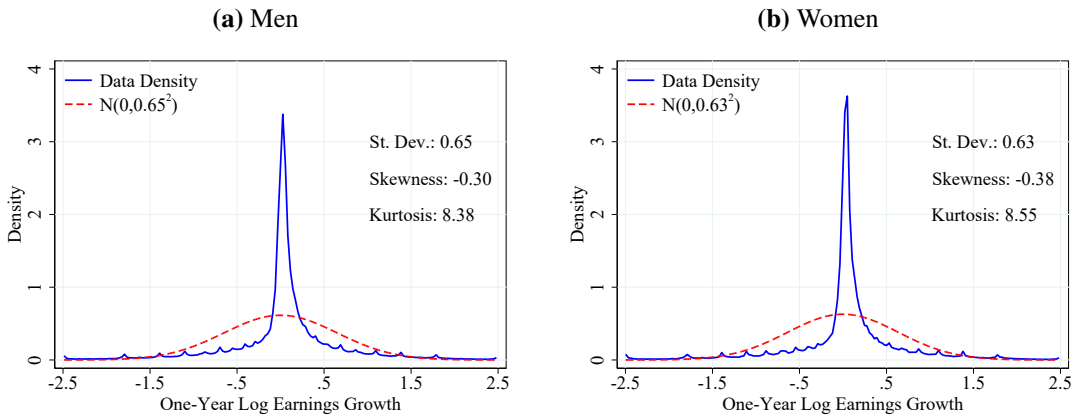
Note: Based on authors' calculations with data from IMSS. The figure shows the log-density of the distribution of one-year earnings changes for men (a) and women (b) in 2005.

Figure B.13: EMPIRICAL LOG-DENSITIES OF FIVE-YEAR EARNINGS CHANGES



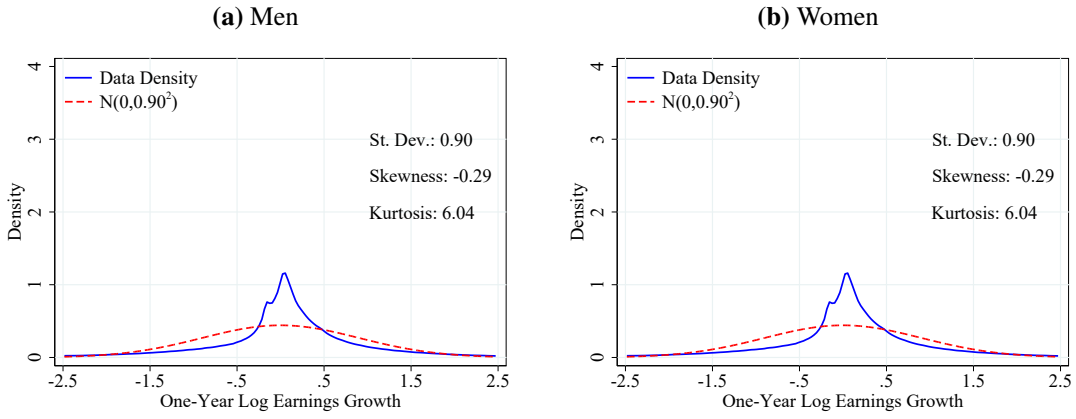
Note: Based on authors' calculations with data from IMSS. The figure shows the log-density of the distribution of five-year earnings changes for men (a) and women (b) in 2005.

Figure B.14: EMPIRICAL LOG-DENSITIES OF ONE-YEAR EARNINGS CHANGES



Note: Based on authors' calculations with data from IMSS. The figure shows the log-density of the distribution of one-year earnings changes for men (a) and women (b) in 2010.

Figure B.15: EMPIRICAL LOG-DENSITIES OF FIVE-YEAR EARNINGS CHANGES



Note: Based on authors' calculations with data from IMSS. The figure shows the log-density of the distribution of five-year earnings changes for men (a) and women (b) in 2010.

C Appendix: Additional Results from the Household Survey

In this appendix we provide more details regarding the comparison exercise discussed in section 4.1. In particular, we analyze more in depth the issue of non-response in the ENOE —the household survey we used— and we present additional results based exclusively on survey data for informal workers and the whole workforce.

Non-response is a well-known drawback of survey data and we believe that it is likely to be the reason why the characterization of the upper part of the distribution of log earnings differs significantly between our administrative and survey data. To illustrate how this issue affects the ENOE data we focus on non-response limited to a specific question in the survey regarding earnings by considering only individuals who report "invalid" earnings. That is, respondents who declared to have remunerated employment and have worked a positive number of hours, but decided to not provide information about their earnings.

Table C.1 presents the average characteristics of individuals in the survey with both valid and invalid earnings and tests for their differences.[38] The main messages of this table are that: (i) the number of individuals with invalid earnings has been significantly growing over time; (ii) based on observable characteristics, individuals with valid earnings are (statistically) different than individuals with invalid earnings; (iii) the characteristics that differentiate these two groups of individuals have been changing over time. Regarding this last point, figure C.1 shows that the evolution of non-response regarding earnings has not only significantly increased over time, but that it has become more prevalent among highly educated workers who live in cities, are employed in the formal sector, and have a full-time job. As these characteristics are usually associated with higher earnings, we conclude that higher earners are those who more frequently choose to withhold information about their income and that they have increasingly chosen to do so. This points to the fact that the ENOE may be particularly inadequate for providing an accurate picture of the top percentiles of the earnings distribution and that we should be cautious when interpreting the statistics that use information from these percentiles.

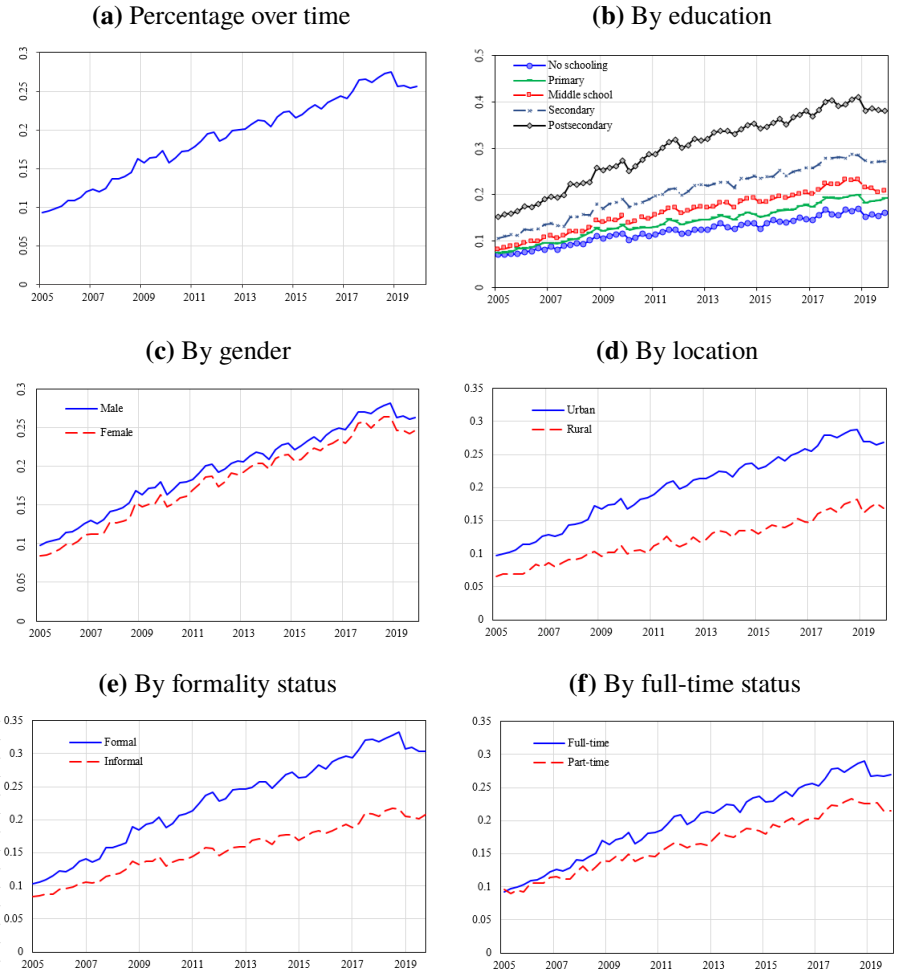
[38] The information comes from the third quarter of selected years.

Table C.1: Differences in means in observable characteristics

Characteristics	2005			2012			2019			Difference in invalid across periods		
	Valid	Invalid	Difference	Valid	Invalid	Difference	Valid	Invalid	Difference	2005–2012	2012–2019	2005–2019
Age	36.86	40.25	-3.40**	37.80	40.80	–2.99**	39.07	41.03	–1.96**	–0.54**	–0.23	–0.77**
			(0.123)			(0.091)			(0.081)	(0.144)	(0.106)	(0.136)
Woman	0.37	0.33	0.04**	0.39	0.37	0.02**	0.40	0.38	0.02**	–0.04**	–0.01**	–0.05**
			(0.004)			(0.003)			(0.002)	(0.005)	(0.004)	(0.005)
Minimum wage stratum	2.96	2.64	0.32**	2.84	2.57	0.27**	2.40	2.28	0.12**	0.07**	0.29**	0.36**
			(0.016)			(0.011)			(0.009)	(0.018)	(0.013)	(0.016)
Formal	0.48	0.56	–0.09**	0.44	0.58	–0.15**	0.47	0.62	–0.15**	–0.02**	–0.03**	–0.05**
			(0.004)			(0.003)			(0.003)	(0.005)	(0.004)	(0.005)
Rural	0.14	0.09	0.04**	0.16	0.09	0.07**	0.13	0.09	0.05**	0.00	0.00	0.00
			(0.003)			(0.002)			(0.002)	(0.003)	(0.002)	(0.003)
Full-time	0.79	0.83	–0.04**	0.76	0.83	–0.07**	0.77	0.84	–0.07	0.00	–0.01**	–0.01**
			(0.004)			(0.003)			(0.002)	(0.004)	(0.003)	(0.003)
No schooling completed	0.18	0.12	0.05**	0.13	0.08	0.06**	0.09	0.05	0.04**	0.05**	0.03**	0.08**
			(0.003)			(0.002)			(0.002)	(0.003)	(0.002)	(0.002)
Primary school	0.23	0.17	0.06**	0.20	0.14	0.07**	0.17	0.11	0.05**	0.04**	0.02**	0.06**
			(0.004)			(0.003)			(0.002)	(0.004)	(0.003)	(0.003)
Middle school	0.33	0.29	0.04**	0.35	0.29	0.06**	0.37	0.28	0.09**	0.00	0.01**	0.01**
			(0.004)			(0.003)			(0.003)	(0.005)	(0.003)	(0.004)
Secondary school	0.10	0.12	–0.02**	0.14	0.16	–0.02**	0.18	0.19	–0.02**	–0.04**	–0.03**	–0.07**
			(0.003)			(0.002)			(0.002)	(0.004)	(0.003)	(0.004)
Postsecondary school	0.16	0.30	–0.14**	0.17	0.34	–0.17**	0.20	0.37	–0.17**	–0.04**	–0.03**	–0.07**
			(0.003)			(0.003)			(0.002)	(0.005)	(0.004)	(0.005)
N. of Observations	132,849	13,977		120,491	29,344		121,850	40,451				

Note: Based on authors' calculations with data from ENOE for the third quarter of 2005, 2012 and 2019. The table reports the means of observable characteristics of individuals whose response to the household survey question about earnings are valid or invalid, and the significance of the difference in these means. The sample includes salaried workers 12 years or older. Standard errors are in parentheses. Stars indicate significance levels (**$p < 0.05$). The table is an updated version of Cuadro 1 in Campos Vázquez [2013].

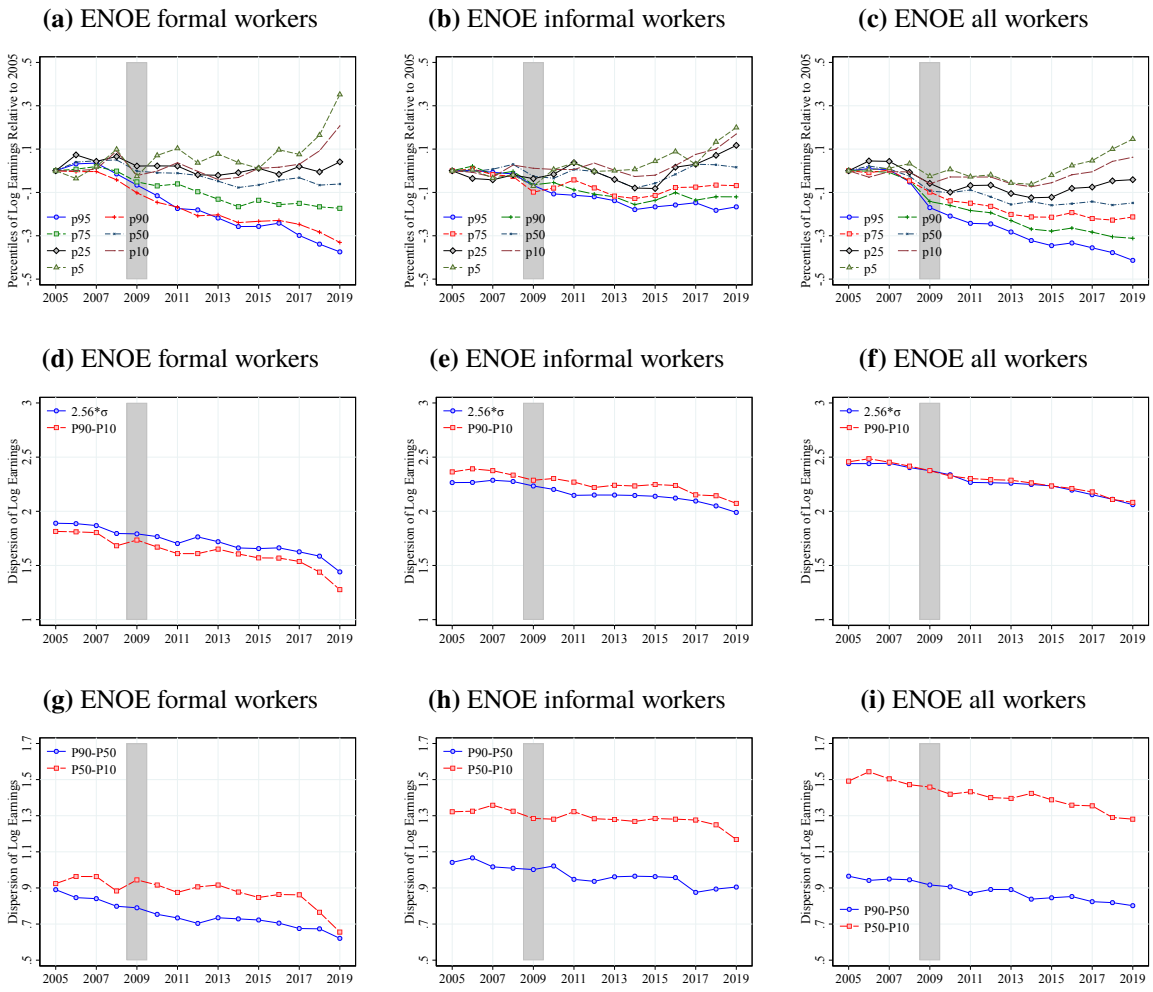
Figure C.1: EVOLUTION OF THE PERCENTAGE OF WORKERS WHO DO NOT REPORT EARNINGS BY SOCIODEMOGRAPHIC GROUPS



Note: Based on authors' calculations with data from ENOE. The figure plots the evolution of the percentage of individuals in the household survey with invalid earnings: (a) over time; (b) over time by educational levels; (c) over time by gender; e) over time by location; (e) over time by formality status; (f) over time by full-time status.

Figures C.3–C.4 are analogous to figures 13–15 presented and discussed in section 4.1. The difference is that here we show the evolution of the percentiles of log earnings and measures of inequality for formal and informal workers, and for the whole pool of workers using exclusively information from the household survey. We find comparable trends indicating that, with the necessary caveats that we have already discussed, the statistics we calculate for formal workers with both administrative and household survey data are relevant and are also relatively in line with those calculated for other categories of workers that are not present in the administrative records.

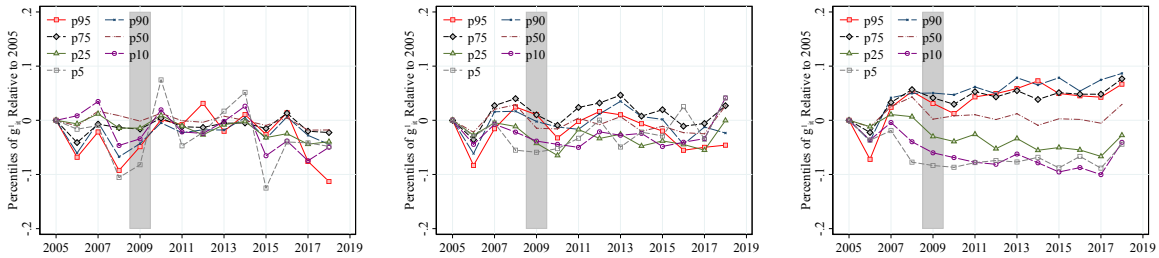
Figure C.2: Comparison between subsamples in the survey data: distribution of log real earnings



Note: Based on authors' calculations with data from ENOE. Using a sample from ENOE constructed to match the IMSS sample as closely as possible, this figure plots against time the following statistics of the distribution of log earnings: (a) ENOE formal workers: P5, P10, P25, P50, P75, P90, P95; (b) ENOE informal workers: P5, P10, P25, P50, P75, P90, P95; (c) ENOE all workers: P5, P10, P25, P50, P75, P90, P95; (d) ENOE formal workers: P90–P10 and 2.56*σ; (e) ENOE informal workers: P90–P10 and 2.56*σ; (f) ENOE all workers: P90–P10 and 2.56*σ; (g) ENOE formal workers: P90–P50 and P50–P10; (h) ENOE informal workers: P90–P50 and P50–P10; (i) ENOE all workers: P90–P50 and P50–P10. Shaded areas are recessions.

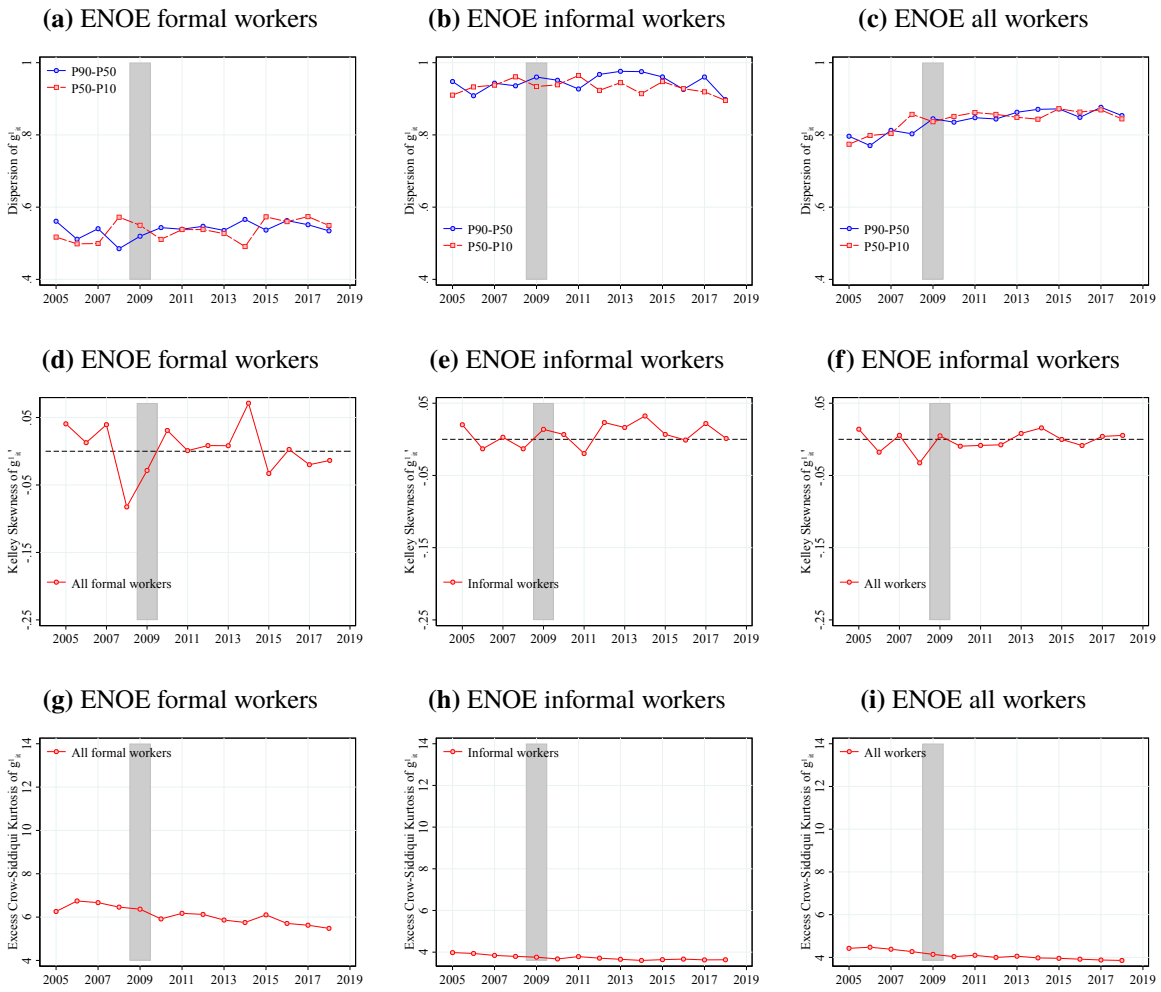
Figure C.3: Comparison between between subsamples in the survey data: evolution of the percentiles of the distribution of one-year log earnings changes

(a) ENOE formal workers **(b)** ENOE informal workers **(c)** ENOE all workers



Note: Based on authors' calculations with data from ENOE. Using a sample from ENOE constructed to match the IMSS sample as closely as possible, this figure plots against time the following statistics of the distribution of one-year log earnings changes: (a) ENOE formal workers: P5, P10, P25, P50, P75, P90, P95, P99, P99.9; (b) ENOE informal workers: P5, P10, P25, P50, P75, P90, P95; (c) ENOE formal workers: P5, P10, P25, P50, P75, P90, P95, P99, P99.9. All percentiles are normalized to 0 in 2005, the first available year. P99 and P99.9 are omitted because, due to the lack of a sufficient number of observations, they are too noisy to be informative. Shaded areas are recessions.

Figure C.4: Comparison between between subsamples in the survey data: measures of dispersion, symmetry, and peakdness of the distribution of one-year log earnings changes



Note: Based on authors' calculations with data from ENOE. Using a sample from ENOE constructed to match the IMSS sample as closely as possible, this figure plots against time the following statistics of the distribution of one-year log earnings changes: (a) ENOE formal workers: P90–P50 and P50–P10; (b) ENOE informal workers: P90–P10 and P50–P10; (c) ENOE all workers: P90–P50 and P50–P10; (d) ENOE formal workers: Kelly skewness; (e) ENOE informal workers: Kelly skewness; (f) ENOE all workers: Kelly skewness; (g) ENOE formal workers: Excess Crow-Siddiqui kurtosis; (h) ENOE informal workers: Excess Crow-Siddiqui kurtosis; (i) ENOE all workers: Excess Crow-Siddiqui kurtosis. Shaded areas are recessions.

D Appendix: Additional Results for Transitions In and Out of Formal Employment

This appendix includes additional information and results that complement those of section 4.2. We present the regression output for the regressions in (4.1) and (4.2) together with a graphical representation of the results of a wide set of robustness exercises we perform. All these exercises confirm that the main results of the analysis in section 4.2 are robust to different specifications.

Table D.1: Estimates of wages trajectories for workers who exit and re-enter formal employment

Dependent variable: Log wages								
Independent variables:	Men		Women		Independent variables:	Men		Women	
$t = -2$	0.062***	(0.007)	0.034***	(0.005)					
$t = -1$	0.047***	(0.007)	0.032***	(0.004)					
$t = -0$	0.000	(·)	0.000	(·)					
$t = 1$	–0.154***	(0.011)	–0.146***	(0.008)					
$t = 2$	–0.081***	(0.009)	–0.063***	(0.008)					
$t = 3$	–0.045***	(0.009)	–0.013	(0.008)					
$k = 1$	0.000	(·)	0.000	(·)					
$k = 2$	–0.019***	(0.007)	–0.035***	(0.006)					
$k = 3$	–0.030***	(0.008)	–0.047***	(0.010)					
$k = 4$	–0.048***	(0.011)	–0.047***	(0.011)					
$k = 5$	–0.047***	(0.014)	–0.060***	(0.013)					
$k = 6$	–0.040**	(0.016)	–0.080***	(0.017)					
$k = 7$	–0.051**	(0.022)	–0.085***	(0.021)					
$k = 8$	–0.097***	(0.021)	–0.105***	(0.018)					
$k = 9$	–0.008	(0.025)	–0.102***	(0.021)					
$t = -2 \times k = 1$	0.000	(·)	0.000	(·)	$t = 1 \times k = 1$	0.000	(·)	0.000	(·)
$t = -2 \times k = 2$	–0.001	(0.008)	–0.004	(0.006)	$t = 1 \times k = 2$	0.001	(0.008)	0.006	(0.009)
$t = -2 \times k = 3$	0.003	(0.008)	–0.002	(0.010)	$t = 1 \times k = 3$	–0.011	(0.009)	–0.009	(0.011)
$t = -2 \times k = 4$	–0.000	(0.012)	–0.008	(0.012)	$t = 1 \times k = 4$	–0.020	(0.013)	–0.028*	(0.014)
$t = -2 \times k = 5$	–0.013	(0.016)	–0.005	(0.013)	$t = 1 \times k = 5$	–0.031*	(0.016)	–0.024	(0.015)
$t = -2 \times k = 6$	–0.008	(0.016)	0.009	(0.011)	$t = 1 \times k = 6$	–0.049***	(0.018)	–0.001	(0.021)
$t = -2 \times k = 7$	0.005	(0.021)	–0.003	(0.024)	$t = 1 \times k = 7$	–0.061**	(0.027)	–0.031	(0.023)
$t = -2 \times k = 8$	0.038	(0.026)	0.030	(0.018)	$t = 1 \times k = 8$	–0.004	(0.027)	0.045	(0.030)
$t = -2 \times k = 9$	–0.029	(0.029)	0.026	(0.021)	$t = 1 \times k = 9$	–0.070*	(0.039)	–0.029	(0.036)
$t = -1 \times k = 1$	0.000	(·)	0.000	(·)	$t = 2 \times k = 1$	0.000	(·)	0.000	(·)
$t = -1 \times k = 2$	–0.002	(0.007)	0.000	(0.006)	$t = 2 \times k = 2$	–0.005	(0.008)	0.004	(0.009)
$t = -1 \times k = 3$	–0.002	(0.008)	–0.004	(0.009)	$t = 2 \times k = 3$	–0.013	(0.011)	–0.014	(0.011)
$t = -1 \times k = 4$	–0.001	(0.011)	–0.001	(0.010)	$t = 2 \times k = 4$	–0.024**	(0.012)	–0.028**	(0.013)
$t = -1 \times k = 5$	–0.005	(0.013)	–0.009	(0.013)	$t = 2 \times k = 5$	–0.040**	(0.017)	–0.026*	(0.014)
$t = -1 \times k = 6$	–0.000	(0.015)	0.006	(0.015)	$t = 2 \times k = 6$	–0.058***	(0.020)	–0.010	(0.021)
$t = -1 \times k = 7$	0.000	(0.021)	–0.010	(0.021)	$t = 2 \times k = 7$	–0.067**	(0.027)	–0.039*	(0.023)
$t = -1 \times k = 8$	0.030	(0.024)	0.018	(0.021)	$t = 2 \times k = 8$	–0.002	(0.033)	0.025	(0.032)
$t = -1 \times k = 9$	–0.020	(0.028)	0.010	(0.026)	$t = 2 \times k = 9$	–0.079**	(0.039)	–0.048	(0.036)
$t = 0 \times k = 1$	0.000	(·)	0.000	(·)	$t = 3 \times k = 1$	0.000	(·)	0.000	(·)
$t = 0 \times k = 2$	0.000	(·)	0.000	(·)	$t = 3 \times k = 2$	–0.003	(0.008)	0.00	(0.008)
$t = 0 \times k = 3$	0.000	(·)	0.000	(·)	$t = 3 \times k = 3$	–0.019*	(0.010)	–0.014	(0.010)
$t = 0 \times k = 4$	0.000	(·)	0.000	(·)	$t = 3 \times k = 4$	–0.016	(0.012)	–0.039***	(0.012)
$t = 0 \times k = 5$	0.000	(·)	0.000	(·)	$t = 3 \times k = 5$	–0.032*	(0.018)	–0.020	(0.015)
$t = 0 \times k = 6$	0.000	(·)	0.000	(·)	$t = 3 \times k = 6$	–0.051**	(0.020)	–0.001	(0.022)
$t = 0 \times k = 7$	0.000	(·)	0.000	(·)	$t = 3 \times k = 7$	–0.083***	(0.028)	–0.037	(0.024)
$t = 0 \times k = 8$	0.000	(·)	0.000	(·)	$t = 3 \times k = 8$	–0.008	(0.032)	0.027	(0.026)
$t = 0 \times k = 9$	0.000	(·)	0.000	(·)	$t = 3 \times k = 9$	–0.090***	(0.027)	–0.025	(0.028)
Constant	8.164***	(0.030)	8.230***	(0.034)					
N. of Observations	682,248 (Men)		389,856 (Women)						

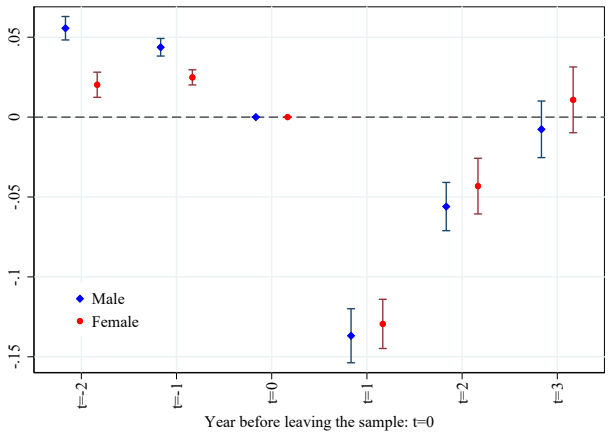
Note: Based on authors' estimates with data from IMSS. The table reports estimates of the coefficients from equation (4.1). The specification includes sector of economic activity, state, and year fixed effects. Standard errors (in parentheses) are clustered at the worker and sector-year levels. Stars indicate significance levels (*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$).

Table D.2: Estimates of wages trajectories: treatment vs control group with a 3-year window

Dependent variable: Log wages				
Independent variables:	Men		Women	
$t = -2$	–0.018***	(0.004)	–0.014***	(0.003)
$t = -1$	–0.011***	(0.003)	–0.007***	(0.002)
$t = 0$	0.000	(·)	0.000	(·)
$t = 1$	–0.001	(0.006)	–0.020***	(0.005)
$t = 2$	0.005	(0.007)	0.025***	(0.006)
$t = 3$	0.008	(0.008)	0.0300***	(0.007)
treated = 1	–0.351***	(0.012)	–0.393***	(0.013)
$t = -2 \times$ treated = 1	0.102***	(0.008)	0.078***	(0.011)
$t = -1 \times$ treated = 1	0.067***	(0.006)	0.053***	(0.007)
$t = 0 \times$ treated = 1	0.000	(·)	0.000	(·)
$t = 1 \times$ treated = 1	–0.189***	(0.016)	–0.222***	(0.016)
$t = 2 \times$ treated = 1	–0..131***	(0.016)	–0.154***	(0.018)
$t = 3 \times$ treated = 1	–0.101***	(0.016)	–0.118***	(0.019)
Constant	8.306***	(0.036)	8.463***	(0.036)
N. of Observations	1,705,014		1,029,924	

Note: Based on authors' estimates with data from IMSS. The table reports estimates of the coefficient $\beta_{\tau}^{\text{treated}}$ from equation (4.2). The specification includes sector of economic activity, state, and year fixed effects. Standard errors (in parentheses) are clustered at the worker and sector-year levels. Stars indicate significance levels (*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$).

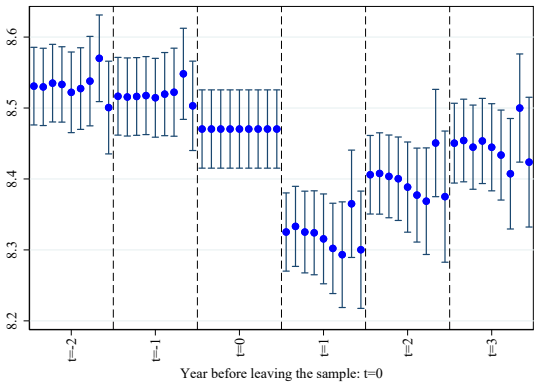
Figure D.1: Estimates of wages trajectories (log differences) of workers who exit and re-enter formal employment adding worker fixed effects



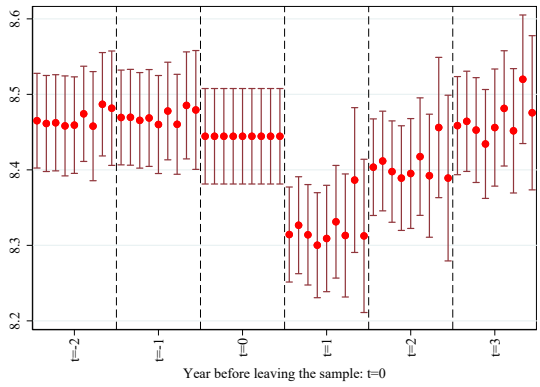
Note: Based on authors' estimates with data from IMSS. The figure plots differences of log wages obtained by estimating equation (4.1) with worker fixed effects using a subsample of workers with only two spells of formal employment and including worker fixed effects in the specification. Markers for men and women are positioned to the left and right, respectively, of each event year t. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates.

Figure D.2: Estimates of wages trajectories (levels) of workers who exit and re-enter formal employment including worker fixed effects

(a) Men

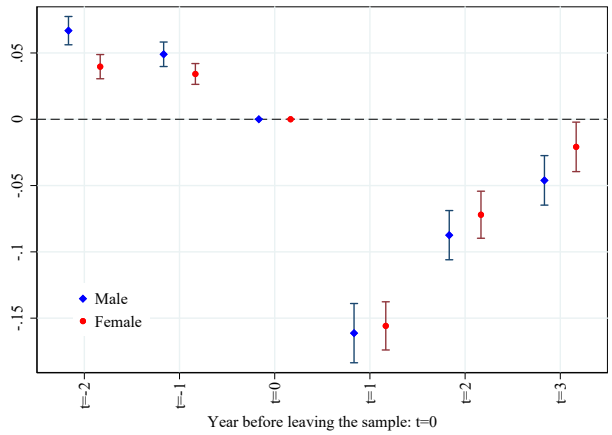


(b) Women



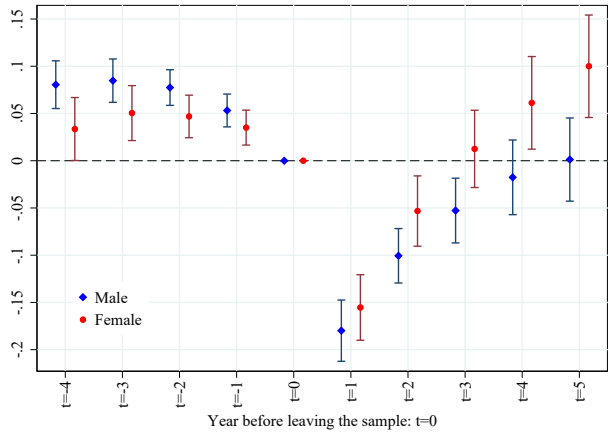
Note: Based on authors' estimates with data from IMSS. The figure plots the conditional means of log wages using the estimated coefficients from equation (4.1) where worker fixed effects are added. The coefficients β_κ and β^κ_τ are omitted in this estimation as they are collinear with worker fixed effects. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates. Standard errors and confidence intervals are obtained with the delta method.

Figure D.3: Estimates of wages trajectories (log differences) of workers who exit and re-enter formal employment including cohort-year fixed effects



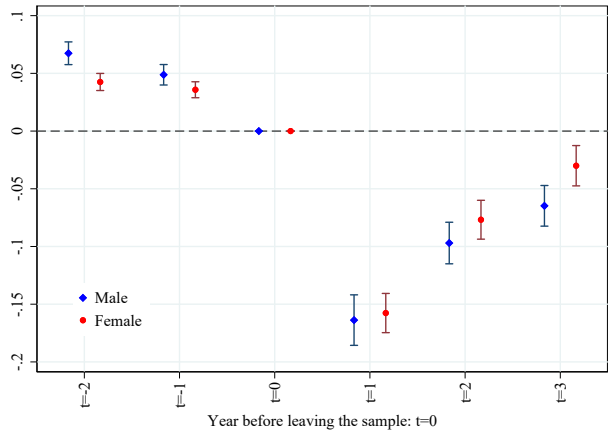
Note: Based on authors' estimates with data from IMSS. The figure plots differences of log wages obtained by estimating equation (4.1) using a subsample of workers with only two spells of formal employment and including cohort-year fixed effects. A cohort c is defined as the cohort of workers who turned 18 in year c. The coefficients β_κ and β_τ^κ are omitted in this estimation as they are collinear with cohort-year fixed effects.Markers for men and women are positioned to the left and right, respectively, of each event year t. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates.

Figure D.4: Estimates of wages trajectories (log differences) of workers who exit and re-enter formal employment with a 5-year event window



Note: Based on authors' estimates with data from IMSS. The figure plots differences of log wages obtained by estimating equation (4.1) using a subsample of workers with only two spells of formal employment and widening the event window to 5 years. Markers for men and women are positioned to the left and right, respectively, of each event year t. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates.

Figure D.5: Estimates of wages trajectories (log differences) of workers who exit and re-enter formal employment with at least 2 spells of formal employment



Note: Based on authors' estimates with data from IMSS. The figure plots differences of log wages obtained by estimating equation (4.1) using a subsample of workers with at least two spells of formal employment. For workers with more than two spells, only the first two are considered. Markers for men and women are positioned to the left and right, respectively, of each event year t. Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates.

As a final robustness check, we assess whether our baseline results from section 4.2 could be driven by the pre-exit and post re-entry wage trajectories of workers whose first spell of formal employment came to an end due to the 2009 financial crisis. We address this concern by estimating the following alternative specification:

$$\ln(w_{it}) = \sum_{\tau=-2}^{3} \beta_{\tau}^{\text{crisis}} \mathbb{I}_{\tau} \mathbb{I}_{i,\text{crisis}} + \gamma_g \mathbb{I}_g + \alpha_e + \alpha_s + \alpha_t + \varepsilon_{it} \tag{D.1}$$

In this case, $\mathbb{I}_{i,\text{crisis}}$ is an indicator variable that equals 1 if the last year we observe the worker in the database before exit is 2008 or 2009. The coefficients $\beta_{\tau}^{\text{crisis}}$ capture the average wage in every year of the event window for workers that exited during the financial crisis as compared to the average wages of those who left in any other year and their estimates are shown in figure D.6. For both genders, we find that one year before exit occurred, workers who left during the financial crisis had, on average, slightly higher wages than those who left in any other year. This difference is statistically significant for men and marginally significant for women. In contrast, the average wage difference among these two groups of workers is not statistically significant upon re-entry. Hence, we conclude that the wage patterns documented with our benchmark specification are a general feature of the transitions out and back into formal employment and do not seem to be driven by the specific exit/re-entry that occurred during the large shock of the financial crisis.

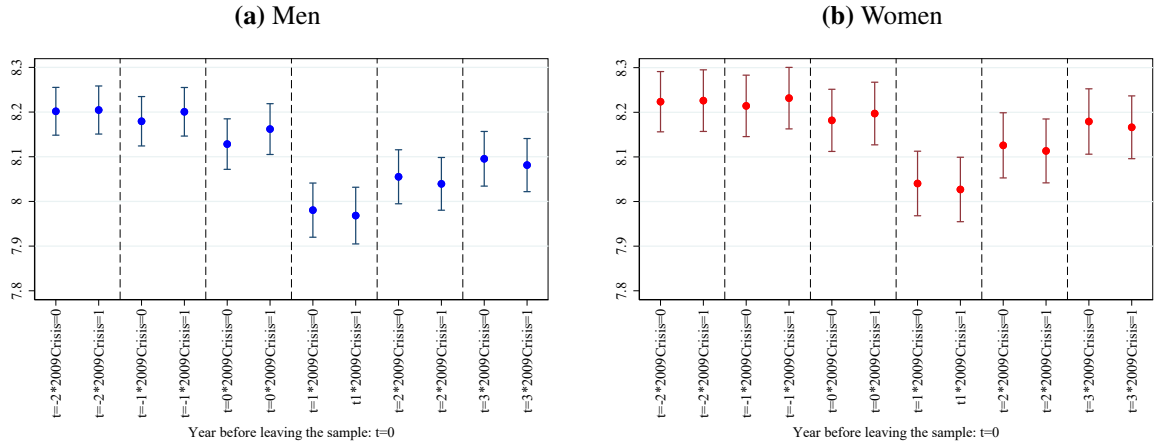
E Appendix: Additional Results for Early Exposure to Informality and Future Earnings

To correct for potential endogeneity, we instrument the formal status of a worker's first job in (4.3) with local labor market informality rates for workers 30 years old and younger. Table E.1 reports the results of a battery of tests that corroborate the validity of our instruments of choice.

Using data from ENE and ENOE, figure E.1 shows the distribution of informality rates in each Mexican state for the period 1995 to 2019 demonstrating that there is a lot of cross-state heterogeneity in terms of the mean and the dispersion of informality rates.

Since information on local labor market informality rates is available only from 1995 onward, the panel that we build from the MOTRAL has to be constrained to include only workers who entered the labor market for the first time in or after 1995. This means that roughly 57% of the observations has to be excluded from the estimation of equation (4.3).

Figure D.6: ESTIMATES OF WAGES TRAJECTORIES OF WORKERS WHO LEFT FORMAL EMPLOYMENT DURING THE 2009 FINANCIAL CRISIS



Note: Based on authors' estimates with data from IMSS. The figure plots the conditional means of log wages computed as $\mathbb{E}[\ln w_{it} | X = X_0, \tau = t, \mathbb{I}_{i,\text{crisis}} = 1] = \hat{\beta}_{\tau}^{\text{crisis}}$ using the estimated coefficients from equation (D.1). Standard errors are clustered at the worker and sector-year levels and 95% confidence intervals are plotted together with point estimates.

Figure E.2 depicts the distribution of year of first entry into the labor market for the whole panel.

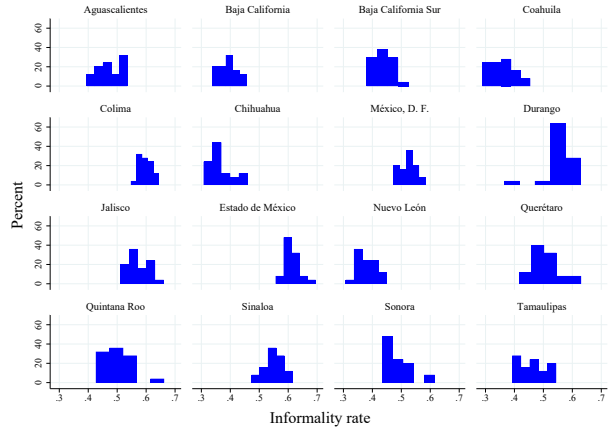
Table E.1: Validity of the instruments used in the first-stage regression

	Low-educated			*High-educated*		
	Men	Women	All	Men	Women	All
Underidentification test						
Kleibergen-Paap rk LM statistic	146.70	53.41	186.27	70.99	31.37	102.84
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Weak identification test						
Cragg-Donald Wald F statistic	69.01	26.35	92.83	35.71	15.87	52.30
Weak-instrument-robust inference						
Anderson-Rubin Wald test	26.57	33.06	64.78	30.82	64.84	57.11
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Stock-Wright LM S statistic	29.38	31.51	64.92	31.41	61.66	56.41
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

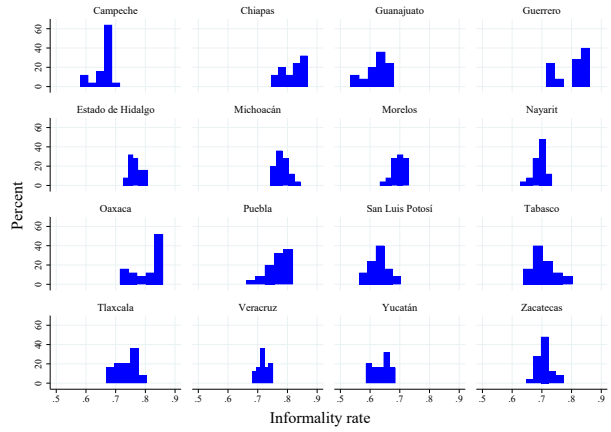
Note: *Underidentification test* — H_0: the matrix of reduced form coefficients has rank 1 (underidentified); H_1: the matrix has rank 2 (identified). The Kleibergen-Paap rk LM statistic has a chi-square asymptotic distribution with 1 degree of freedom. *Weak identification test* — H_0: the equation is weakly identified. The critical values for the Cragg-Donald Wald F statistic for weak instrument based on LIML with 10% maximal IV size is 7.03 for the case of 2 endogenous regressors and 2 excluded instruments (see Stock, Yogo, et al. [2005]). *Weak-instrument-robust inference* — Tests of joint significance of endogenous regressors in the main equation. H_0: $\beta_3 = \beta_4 = 0$ and orthogonality conditions are valid. The Anderson-Rubin Wald test and Stock-Wright LM S statistics have both a chi-square asymptotic distribution with 2 degrees of freedom. p-values are reported in parenthesis.

Figure E.1: Distribution of Informality Rates per State

(a) States with low informality rate

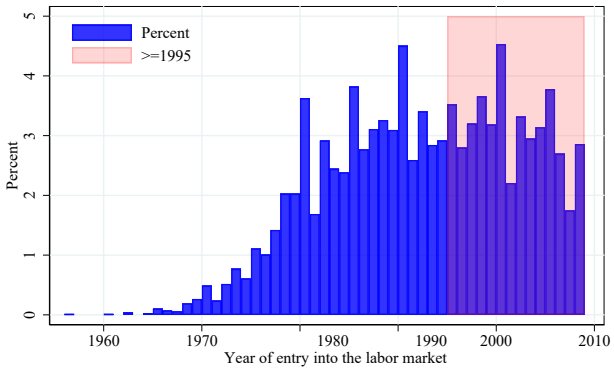


(b) States with high informality rate



Note: Based on authors' estimates with data from ENE and ENOE. The figure plots the average rate of informality for the Mexican states with (a) low levels of informality and (b) high levels of informality during the period 1995–2019.

Figure E.2: Distribution of the Year of Entry into the Labor Market



Note: Based on authors' estimates with data from MOTRAL. The figure plots the frequency of year of entry for all the individuals present in the panel we constructed from the two rounds of the MOTRAL and highlights the observations that were actually included in the analysis presented in section 4.3.