Recovering the counterfactual wage distribution with selective return migration

Article

Accepted Version

Recovering the Counterfactual Wage Distribution with Selective Return Migration*

Costanza Biavaschi†

Forthcoming in *Labour Economics*
doi: 10.1016/j.labeco.2015.12.001

Abstract

This paper recovers the distribution of wages for Mexican-born workers living in the U.S. if no return migration of Mexican-born workers occurred. Because migrants self-select in the decision to return, the overarching problem addressed by this study is the use of an estimator that also accounts for selection on unobservables. I find that Mexican returnees are middle- to high-wage earners at all levels of educational attainment. Taking into account self-selection in return migration, wages would be approximately 7.7% higher at the median and 4.5% higher at the mean. Owing to positive self-selection, the immigrant-native wage gap would, therefore, partially close if there was no return migration.

JEL Classification Codes: J61, F22
Keywords: return migration, self-selection, assimilation, U.S.-Mexico migration.

1 Introduction

Migrants' self-selection is a core issue in labor economics. If migrants are rational actors optimally choosing their residence location, any observed outcome for this group will be endogenous to the original migration decision. Hence, to understand migrants' outcomes it is necessary to understand the nature of their selection. Yet, the literature has primarily viewed migration as permanent, when in fact individual migration is often of a temporary nature.

The recognition that migration is a dynamic process has more recently encouraged scholars to understand its drivers (Dustmann, 2003) and its consequences in terms of migrant selectivity (Borjas and Bratsberg, 1996; Dustmann and Weiss, 2007). How do returnees compare with those who

*Acknowledgements: I would like to thank Corrado Giulietti, Roger Klein, Carolyn Moehling, Anne Piehl, Helena Skyt Nielsen, two anonymous reviewers as well as conference and seminar participants at IZA, NHH and at the 4th AFD-World Bank International Migration and Development Conference for their helpful comments. All errors are mine.

†Department of Economics, University of Reading, Whiteknights, Reading, RG6 6AA, United Kingdom. Email: c.biavaschi@reading.ac.uk

permanently settle abroad? Answering this question is consequential for several lines of research. From the destination country perspective, a vast literature has attempted to measure the economic assimilation of immigrants with natives (see seminal work by Chiswick (1978) and papers by Borjas (1985); LaLonde and Topel (1992); Borjas (1994) among others). If selectivity in return migration is not considered, however, the economic progress of immigrants will be over- or underestimated depending on the nature of this selection (Hu, 2000; Lubotsky, 2007). From the source country perspective, return migration may mitigate the brain drain through acquisition of skills used at home (Dustmann et al., 2011). Hence, return migration may help to foster growth in the source country through an expansion in its human capital stock (Dos Santos and Postel-Vinay, 2003). Taking into account selectivity in return migration urges scholars to reconsider how they measure the effects of migration on both immigrants and natives, as well as on both the sending and receiving regions.

Building on the previous literature that often analyzes how returnees' average earnings differ from those of stayers, this paper combines data derived from U.S. and Mexican censuses to estimate the wage *distribution* of Mexican-born immigrants in the U.S. under two conditions, namely – with and without return migration. This approach enables answering two key questions: how do returnees compare with stayers and where does return migration have its largest impact on the wage distribution?[1] This paper highlights the consequences for the U.S. if no return migration of Mexican-born workers had occurred between 1995 and 2000, shedding light on a counterfactual scenario that could have occurred if incentives to return were altered based on exogenous variations in economic opportunities in the source or host countries.

The overarching problem of this study is to recover the counterfactual wage density in the presence of selective return migration, when pre-migration earnings are not known. Crucial to the approach adopted is the introduction of an estimation technique that can recover such distribution, taking into account not only the observable differences between stayers and returnees but also self-selection on unobservables. This paper proposes a semiparametric procedure that complements the estimator presented by DiNardo et al. (1996) applied in the migration literature (Butcher and DiNardo, 2002; Chiquiar and Hanson, 2005), which accounts for selection based on observable traits only. The presented estimation method is based on the observation that selection bias disappears for subgroups where nearly all individuals settle permanently in the U.S. This procedure provides an alternative to the use of pre-migration earnings to measure selectivity, as these are often unavailable to the researcher either due to the lack of longitudinal data following stayers and returnees or because the return flows in available surveys are often too small to allow suitable analysis.

Conditioning on observable characteristics, I find that Mexican returnees are middle to high wage earners, consistent with models in which the decision to return hinges on reaching target-earnings levels. Taking self-selection into account, the wages of Mexican born workers in the U.S. would be approximately 7.7% higer at the median and 4.5% higher at the mean. Furthermore, the

[1] I assume throughout that the supply effects of the absence of return migration are negligible. Given the negative yet often small impact of migration on the overall economy, this assumption seems to be reasonable.

return flow has a small effect on immigrant wage inequality: the outflow of immigrants increases dispersion in the lower part of the distribution and decreases it in the upper part. Moreover, selective return migration does not have a constant effect across educational levels: while it increases inequality at low levels of education, it decreases inequality for the highly skilled. These results suggest that when designing optimal migration policies policymakers should consider that selective outmigration might have a greater impact at high levels of human capital. Finally, because at all levels of education the immigrants who leave are the high-wage earners, the immigrant–native wage gap would close slightly if there was no return migration.

The remainder of this paper is organized as follows. Section 2 reviews the literature. Section 3 describes the data. Section 4 presents the estimation technique and Sections 5 and 6 the results, while Sections 7 and 8 show the sensitivity of the results to different specifications and to the assumptions made. Section 9 concludes the paper.

2 Immigration, Return Migration and Self-Selection across the Mexican–U.S. border

Several contributions to the immigration literature have empirically assessed the selection of immigrants from Mexico to the U.S., while the literature on the selection of Mexican return migrants is relatively less developed.

The current debate on immigrant selection has developed from the results of Chiquiar and Hanson's (2005) which contradict the theoretical predictions proposed in Borjas (1987), showing intermediate to positive selection based on the observable characteristics of Mexican immigrants to the U.S. compared with Mexican stayers in Mexico. Yet the finding of positive selection was challenged by a few authors (Ibarraran and Lubotsky, 2007; Fernandez-Huertas Moraga, 2011; McKenzie and Rapoport, 2010; Ambrosini and Peri, 2012; Kaestner and Malamud, 2014), who have drawn scholarly attention to the importance of two key elements in the analysis of the selectivity of migrants. First, it is crucial to use nationally representative data sources that have a longitudinal component capable of capturing the pre-migration earnings of migrants and non-migrants (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012; McKenzie and Rapoport, 2010; Kaestner and Malamud, 2014). Second, researchers must aim to control for the unobservable differences between migrants and non-migrants (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012).

Turning to the selection of returnees, the overall evidence for the U.S. economy suggests that returnees have below average skills. By comparing longitudinal and cross-sectional data, Lubotsky (2007) finds that return migration by low-wage immigrants from the U.S. has systematically led past researchers to overestimate the wage progress of stayers by 10% to 15%. Likewise, Hu (2000) shows a decline in immigrant wage growth once return migration has been taken into account, with such results being weaker for Hispanic workers. Hu (2000) and Lubotsky (2007) both provide interesting insights into the nature of return migration and its impact on the host economy; however,

in their longitudinal datasets returnees are not directly identified and return migration cannot be separated from other sources of panel attrition.[2] Furthermore, their estimation technique is based on the assumption of time invariant unobserved selection.[3]

The previous discussion confirms that self-selection and data availability have limited our understanding of return migration and its consequences. Therefore, in order to fill this gap in the literature, this paper advances an analysis that uses representative data and examines the actual return choices of Mexican migrants based on a dataset that combines data from both U.S. and Mexican censuses. While combining census data to study return migration is not novel and was used in Lacuesta (2010), this study adds to that contribution by controlling for selection on unobservables. The use of two censuses together with the econometric technique proposed allows researchers to distinguish return migration from panel attrition and to treat all those forms of sample selection and heterogeneity that are not simply eliminated by fixed effects estimators in panel data analyses. Furthermore, it provides a full picture of what the U.S. could expect if return migration was zero, owing to changes in either migration policies or migration incentives.

On the methodological side, this paper introduces an estimator for a counterfactual distribution that accounts for sample selection. This technique complements the analysis based on selection on observables (Chiquiar and Hanson, 2005; Ibarraran and Lubotsky, 2007)[4] in order to account for selection on unobservables as well. The proposed estimator is based on the model presented by Heckman (1990), and it extends the estimator proposed by Andrews and Schafgans (1998) to its density equivalent. This method could also be applied to other contexts in order to recover a distribution of outcomes that are truncated and/or when panel data are unavailable.

3 Data

The analysis presented uses the U.S. and Mexican censuses from 2000, available through the International IPUMS website.[5] The Mexican census was conducted in February 2000 by the Instituto Nacional de Estadística Geografía e Informática (INEGI), the Mexican statistical agency. Household heads were asked to list all current members of the household and to also list their place of

[2] In particular, these authors identify non-employment, outmigration, employment in the informal sector, and nonmatch as possible causes of panel attrition.

[3] Further analyses from the Mexican perspective include Lacuesta (2010), Ambrosini and Peri (2012) and Reinhold and Thom (2013) . Lacuesta (2010) and Reinhold and Thom (2013) both provide evidence of selection and skill upgrading for Mexican returnees in Mexico. Lacuesta (2010) argues that return migrants are similar to stayers, suggesting that the 7% wage premium found upon return might actually be caused by the selection of return migrants that were unaccounted for in the analysis. Meanwhile, Reinhold and Thom (2013), using the Mexican Migration Project (which is not a representative sample), estimate the experiences of returnees to the U.S. labor market by correcting for the endogeneity of migration decisions. They find that returnees are negatively selected in terms of unobservable traits, although selection is not significant in their analysis. Finally, Ambrosini and Peri (2012) find preliminary evidence that returnees are positively selected compared with non-migrants and permanent migrants. However, the results on returnees' self-selectivity are based on a very small sample.

[4] Chiquiar and Hanson (2005)'s estimation is in turn based on DiNardo et al. (1996)

[5] Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 6.4 [Machine-readable database]. Minneapolis: University of Minnesota, 2015. See `https://international.ipums.org/international/`, last retrieved on June 2012.

residence in 1995. The data consist of a 10.6% sample of the Mexican population and are stratified geographically by municipality and urban area. The U.S. census was conducted in April 2000 and the available data are a 5% random sample of the original forms. It is a weighted sample with stratification by state. Mexican-born immigrants are defined as individuals born in Mexico who appear in the U.S. census. Mexican-born return migrants appear in the Mexican census and are identified as those individuals who report having resided in the U.S. five years prior to the Mexican census enumeration. The data on the returnees in Mexico and the data on the Mexican-born stayers in the U.S. are then pooled to build a dataset containing all individuals who are or have been in the U.S. For comparison purposes, this study also uses data on a random sample of U.S. native-born workers (n = 103,994).[6]

However, the use of different data sources to identify return migrants is not without limitations. As discussed in Chiquiar and Hanson (2005) and Ibarraran and Lubotsky (2007), the most notable drawbacks relate to changes in education once in the U.S., the misreporting of education in the U.S. census, and undercount due to illegal immigration and short trips. Given that this study focuses on return migration, the possibility of Mexican immigrants having obtained additional schooling after arriving in the U.S. should not be regarded as invalidating since returnees could have made the same choice. Nonetheless, the concern remains that Mexican migrants in the U.S. might overstate their levels of educational attainment (Ibarraran and Lubotsky, 2007). I will further discuss this issue when checking the robustness of the results. The undercount of illegal immigrants and short-term migrants in the U.S. census might indeed constitute a problem, which is discussed in Section 7. Finally, there is a further concern specific to this study: the universe of returnees is much broader than is captured by the Mexican census. If Mexican workers who returned before 1995 systematically differ from those who returned between 1995 and 2000, the conclusions of this paper would not be externally valid. Since no further information is available on workers having been abroad, looking at place of residence in 1995 is the best proxy for return status. [7]

The sample is restricted to men aged between 25 and 55 years, born in Mexico, and in earning employment, resulting in a total sample size of 133,389. Of this number, 120,205 (90%) immigrants stay in the U.S., while 13,184 (10%) are return migrants. This study applies four indicators of educational attainment (*Less than primary school completed, Primary school completed, Secondary*

[6] I have run exploratory analyses to understand whether the use of weighting in each sample affected the conclusions of the paper. Results are comparable with those reported in the analysis, and in particular the selection of individuals with a high probability of staying is unaffected by the weighting procedure. The sampling criteria of the censuses are unlikely to be related with the decision to stay or with the error term in the wage equation. If the reader is concerned about the geographical stratification, the full model estimated in Section 7 should reassure that results are stable once we control for geography. In choosing not to use weight, the results are directly comparable with those in the literature (Chiquiar and Hanson, 2005; Lacuesta, 2010). Additional details on sampling issues are discussed in Section 7.

[7] The 1990s were a decade of radical transformation in the Mexican economy, with the signing of the NAFTA in 1994, the Mexican peso crisis in 1994-1995, and the subsequent period of macroeconomic growth. It is, therefore, possible that changing macroeconomic conditions affected the return migration flow. However, it remains of interest to study the return phenomenon during periods of financial turmoil, as public opinion might have particular sentiments about migrants prolonging their stays in the U.S. Finally, a parametric analysis of selection in 1990 shows a similar pattern of return migration. These results are available upon request In addition, Section 7 further shows how selection patterns persist when looking at return migration in the early 2000s.

school completed, College Degree), while socioeconomic characteristics are represented by indicators of being married (*Married*), having children (*Child*), and having a U.S.-born spouse (*Spouse U.S.-born*) or child (*Child U.S.-born*). The decision to stay in the U.S. is modeled throughout this analysis as a function of these educational and socioeconomic variables. Table C.1 in the Appendix reports the characteristics by return status. [8]

The wage process is modeled according to various specifications. In the first set of regressions the observable characteristics include those regressors used in previous analyses on Mexican-U.S. selection, namely, education, age and family status (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012; McKenzie and Rapoport, 2010; Lacuesta, 2010; Kaestner and Malamud, 2014). Moreover, the indicator of having a U.S.-born spouse is included to capture the constructs of "attachment"and "networks,"which have both been shown to be relevant determinants of migration decisions (McKenzie and Rapoport, 2010; Ambrosini and Peri, 2012). Having a U.S.-born child is also included in the model for the decision to stay, yet is excluded from the wage equation. To summarize the effects of these characteristics on the wage distribution, Figure 1 applies the methodology developed in DiNardo et al. (1996) to show the actual distribution of wages for Mexican-born workers in the U.S. and the distribution of wages that would have occurred if U.S. stayers shared the observable characteristics of returnees and were paid according to U.S. skill prices.

[FIGURE 1 HERE]

Figure 1 illustrates that the counterfactual distribution is shifted to the left compared with the actual distribution observed in the U.S.: therefore, returnees appear to be negatively selected in terms of observable characteristics. Consequently, based on observable traits, the figure suggests that returnees are drawn disproportionately from the bottom part of the wage distribution. Section 4 explains how a counterfactual distribution can be estimated if all returnees had stayed, by accounting for both observable and unobservable traits. The remainder of the paper subsequently compares the counterfactual results with the descriptive analysis presented in this section.

4 The Model and the Estimation Strategy

The research question answered in this study requires recovery of the wage distribution for all Mexican-born men who have been in the U.S., despite wages only being observed for Mexican-born immigrants who are currently residing in the country. Let S_i be an indicator of whether or not

[8] Table C.1 in the Appendix reports also the average characteristics of other variables used in the analysis. Experience in the U.S. is represented by indicators of length of stay between 0 and 5 years, 5 to 10 years, 10 to 20 years, 30 to 40 years, and more than 40 years (*Years in the U.S.*). The limited information collected by the Mexican census about returnees' experiences abroad means that how long these workers stayed in the U.S. before returning to Mexico is unknown. Regional labor market characteristics are represented by indicators of residence location in four regions: *West, Northeast, Midwest, South.* Fourteen industry variables are also reported. The table further reports the average wages of U.S. stayers. The wage variable is constructed as wage and salary income divided by hours of work. To avoid division bias (Borjas, 1980), earnings were also used as the dependent variable, without changes to the conclusions of the paper. The average wage in the U.S. for returnees is unobserved.

individual i decides to stay in the U.S. In the following model this decision depends on the net benefits of staying, $(Z_i'\alpha_0 - \epsilon_i)$, being greater than zero.

Let r be the number of returnees and n be the number of stayers. The decision to stay can be represented as:

$$S = \begin{cases} 1 & Z_i'\alpha_0 > \epsilon_i \\ 0 & Z_i'\alpha_0 \leq \epsilon_i \end{cases} \qquad \text{for} \quad i = 1, \ldots, r+n \tag{1}$$

The true wage determination process for a randomly selected Mexican immigrant present in the U.S. takes the form:

$$Y_i^* = X_i'\beta_0 + c_0 + u_i^* \qquad i = 1, \ldots, r+n. \tag{2}$$

In the model, Y_i^* is the log of the hourly wage for Mexican immigrants, and X_i represents the determinants of the log-wage process.

The wage is observed only for those immigrants who stay in the U.S., however. In other words, the observed wage is:

$$Y_i = S_i Y_i^* \qquad i = 1, \ldots, r+n. \tag{3}$$

From the model in equations (1) and (2) it follows that (Y, S_i, X_i, Z_i) are observed random variables. The aim of the estimation is to obtain the distribution of Y_i^*, given that only Y_i is observed. For generality, the remainder of the paper focuses on estimation techniques that are free from distributional assumptions, while a comparison with the parametric model is reported as a robustness check. Using flexible estimators is particularly important whenever the parametric assumptions are not satisfied. It is assumed throughout that $(X_i, Z_i, u_i^*, \epsilon_i)$ are i.i.d and (X_i, Z_i) are exogenous random variables. It should be stressed that u_i^* and ϵ_i are allowed to be correlated. Section 7 discusses these assumptions. The next subsection introduces the estimation strategy.

4.1 Counterfactual Density Estimation

The distribution of Y_i^* in equation (3) corresponds to the distribution of u_i^* up to a location shift represented by the observable characteristics, $(X_i'\beta_0 + c_0)$. Hence, the estimation strategy of the counterfactual wage distribution entails two steps:

1. Recover the distribution of the unobservables, $f(u^*)$
2. Recover the full distribution of wages by shifting $f(u^*)$ by the mean of $(X_i'\beta_0 + c_0)$. At this stage *all* observable differences between stayers and returnees are considered and taken into account.

Because the key challenge and contribution of the paper lies in recovering $f(u^*)$ (step 1), while standard techniques can be adopted to estimate β_0, c_0 (step 2), the following discussion largely

focuses on recovering the distribution of u_i^*.

Let $f(u_i^*)$ be the unknown distribution of u_i^*. By the Law of Total Probability, $f(u_i^*)$ can be written as a weighted sum of the distribution of the error terms in the subsamples of stayers and returnees with weights given by the probability of being in either subsample, i.e.:

$$f(u_i^*|Z_i'\alpha_0) = f(u_i^*|S_i = 1, Z_i'\alpha_0)\Pr(S_i = 1|Z_i'\alpha_0) + f(u_i^*|S_i = 0, Z_i'\alpha_0)\Pr(S_i = 0|Z_i'\alpha_0),$$

I will assume throughout that $f(u_i^*|S = 1, Z_i'\alpha_0) = f(u_i^*|S = 1)$ and similarly $f(u_i^*|Z_i'\alpha_0) = f(u_i^*)$. The analysis is carried out under this assumption for computational speed and expositional purposes. Independence is, however, not necessary. The results are similar when conditioning on particular quantiles of the selection index, as further discussed in Section 8.

This density cannot be directly estimated using the sample wage distribution, as the latter is only observed conditional on the decision to stay. In other words, it is not possible to directly obtain an estimate of $f(u_i^*)$ as no information can be directly extrapolated from the data about the unobservable component of the returnees' wage equation. Otherwise stated, $f(u_i^*|S_i = 0, Z_i'\alpha_0)$ is unknown. However, note that whenever $\Pr(S_i = 1|Z_i'\alpha_0)$ is close to 1, $f(u_i^*|Z_i'\alpha_0) = f(u_i^*) \approx f(u_i^*|S = 1, Z_i'\alpha_0) = f(u_i^*|S_i = 1)$. Intuitively, selection disappears in the limit for individuals for whom $\Pr(S = 1|Z_i'\alpha_0)$ is close to 1, namely for those individuals in a *high probability* set. While it is not possible to know the wage distribution of the returnees, we can still recover the counterfactual distribution out of a subsample for which the likelihood of staying is very high.

This intuition is known as *identification at infinity* (Chamberlain, 1986). Identification at infinity has been advocated by Heckman (1990) to estimate the constant term in semiparametric sample selection models and Andrews and Schafgans (1998) develop an estimator in the spirit of the one proposed by Heckman.[9] Several applications have relied on this identification strategy to recover the constant term in sample selection models, ranging from studies of gender (Schafgans, 2000; Martins, 2001; Mulligan and Rubinstein, 2008; Albrecht et al., 2009; Chzhen and Mumford, 2011) and ethnic wage differentials which correct for labor force participation choices (Schafgans, 1998) , to the estimation of wage differentials for union and non-union members (Lanot and Walker, 1998), to the effects on children's BMI of mothers' labor force participation choices (Liu et al., 2009), to health care utilization differences among insured and uninsured workers (Shen, 2013). In all these examples, the constant is recovered out of a subset of individuals for which selection is unlikely to occur, i.e. groups with fairly high participation rates (older, urban workers with secondary degrees), likely to be unionized (older, semi-skilled workers) or insured (individuals with several co-morbidities working in industries with high insurance rates). As long as the characteristics Z that select the sample are exogenous and observations are i.i.d., focusing on particular groups where selection disappears should deliver a consistent estimate of the quantity of interest. In my context, this idea will be used to recover $f(u^*)$. Applying this strategy, estimation of the distribution of

[9] Schafgans and Zinde-Walsh (2002) prove the asymptotic properties of Heckman's (1990) proposed estimator for the intercept in a sample selection model, while Klein et al. (2015) extend these results to allow for a definition of an high probability set that is data-dependent.

unobservable characteristics is based on a subsample selected on observed characteristics Z, in which nearly all individuals stay in the U.S. To the best of my knowledge, such an identification strategy has not been applied to recover a counterfactual distribution, as is carried out in this paper, in order to allow for selection on unobservables in a counterfactual density estimation.

Let H_i be an indicator that defines whether the observation is in this high-probability set, i.e. let $H_i = 1[\Pr(S_i = 1|Z_i'\alpha_0) > \bar{p}_n)]$. The proposed estimator for $f(u_i^*)$ is:

$$\widehat{f(u_i^*)} = \frac{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{u-u_i^*}{h}\right) S_i H_i}{\sum_{i=1}^{n} S_i H_i}, \tag{4}$$

where $K(\cdot)$ is a kernel density estimator and h is the bandwidth parameter. A Gaussian kernel with optimal bandwidth $h = 1.06\hat{\sigma}N^{-1/5}$(Silverman, 1986) is chosen throughout the paper when reporting the densities of interest. This estimator is simply a kernel density estimator of the random variable u^* over a proportion of observations for which the probability of being in the selected sample is close to 1 at the limit. A Monte Carlo is reported in Appendix A to explain how well this method works.

4.2 Parameter Estimation

To estimate the density in equation (4), unbiased estimates of the parameters in the model (α_0, β_0, c_0) must be obtained in order to construct the residuals, $\hat{u}^*$. To study the S_i choice, I estimate a semi-parametric dichotomous choice model,[10] by applying the estimation method developed by Klein and Spady (1993). This estimator is the semi-parametric equivalent of a standard logit or probit model. In fact, the parameters of interest are estimated by maximizing a log-likelihood function where the probability of staying in the U.S. is a semiparametric expectation function of the single-index $Z_i'\alpha$, rather than a parametric probability in a logit/probit form function of the same index. Therefore, the likelihood function takes the standard following form:

$$\widehat{\ln L} = \sum_{i=1}^{n} S_i \ln(\hat{P}_i(v_i)) + (1 - S_i) \ln(1 - \hat{P}_i(v_i)),$$

where v_i is the selection index $Z_i'\hat{\alpha}$ hereafter. In the construction of the likelihood, some of the observations for which this probability is poorly estimated are trimmed.[11] To estimate $\hat{P}$, I apply the bias correction technique proposed by Klein and Shen (2010) to overcome finite sample performance issues.[12] While the original formulation of Klein and Spady assumes that one of the variables in

[10] On the contrary, DiNardo et al. (1996) choose to adopt a parametric specification for their selection model (hence, their approach is deemed to be semiparametric). For coherence, I estimate all parts of the model without any distributional assumptions. In Section 5, however, I also present parametric estimates for comparison.

[11] Trimming is standard in this literature.

[12] Both in Klein and Spady (1993) and in Klein and Shen (2010), the bandwidth h is set to satisfy $n^{-1/6} < h < n^{-1/8}$. Specifically, I set $h = n^{1/7}$. Additionally, it should be noted that in the estimation of the selection index, the only identified parameter in terms of the original model is the coefficient ratio, i.e., α_j/α_1, with $j = 1 \ldots k$ and where α_1 is the coefficient of the continuous variable, which is normalized to 1.

the index is continuous and with non-zero coefficient to reach identification of the semiparametric expectation, subsequent work in Delgado and Mora (1995) and Klein and Shen (2010) suggests that this assumption is sufficient but not necessary for identification, provided tail conditions on the density for the dependent variable are satisfied. In this application, I have used the variable *Age* as normalization variable. Age takes 31 distinct values, and Monte Carlo simulations available upon request on the performance of the Klein and Spady estimator in my setting show that this variation is sufficient to identify the parameters of interest.

The recovery of $Z_i'\hat{\alpha}$ is useful for two reasons.

First, it is now possible to select those observations in the high probability set, for which selection can be ignored at the limit. Thus, individuals in the high probability set represent those observations in the 95th percentile of $\widehat{\Pr(S_i = 1|Z_i'\hat{\alpha})}$.[13]

Second, the estimation of $(Z_i'\hat{\alpha})$ allows us to obtain unbiased estimates of the outcome equation parameters. In the wage equation, I employ Robinson's (1988) differencing method in order to correct for sample selection and recover unbiased estimates of β_0, as well as the estimator proposed by Heckman (1990) to recover c_0.

Before proceeding to the results, one identification issue must be discussed. At least one variable is needed in the Z_i matrix that does not appear in the X_i matrix. The variable included in the selection process and excluded from the wage process is having a U.S.-born child, which proxies for social attachment to the destination country. Because the idea of attachment to people and institutions in the destination country raises the opportunity cost of returning, this should act as a strong predictor of this choice. However, it is unlikely that the wage process depends on the birthplace location of an individual's children.[14] Consequently, the effect of having a U.S.-born child should not predict an individual's wage, after controlling for attachment and network effects through the U.S.-born spouse indicator and length of stay in the U.S. variable.[15]

4.3 Potentials and Limitations of the Estimation Strategy

Primary assumptions are that $(X_i, Z_i, u_i^*, \epsilon_i)$ is i.i.d and that the regressors are exogenous. These are common assumptions of standard empirical models in the migration literature. In fact, the mean independence of the error term from the explanatory variables in the outcome and selection equations is typical in linear regression models (Lacuesta, 2010; Reinhold and Thom, 2013; Kaestner and Malamud, 2014) and in the non-parametric analyses based on pre-migration earnings (e.g., Ambrosini and Peri, 2012; Fernandez-Huertas Moraga, 2011) as selection is here recovered only if the subdivision into cells is exogenous.[16] These models further conjecture the absence of an

[13] Although this cut point is arbitrary in the paper, results are stable when a different definition of the high probability set is used. Such results are available upon request.

[14] Additional regressions also controlled for the language spoken at home. Having a U.S. child might be related to an individual's wage if English proficiency is enhanced by the presence of a child at home. The results that control for this additional variable do not differ from those presented in the paper, and they are available upon request.

[15] Additional evidence and discussion on the validity of the exclusion restriction can be found in Section 8.

[16] As already mentioned, the variables used here to identify individuals that have a high probability of staying are similar to those used to identify cell-probabilities in other studies.

Ashenfelter dip, and that expectations of migration and return do not influence the individual's behavior before migrating. The above estimator avoids these hypotheses as it does not use pre-migration earnings to measure selection. Nevertheless, it imposes a structure between the outcome and observable traits and a stronger need for identically distributed observations. This assumption is further discussed in Section 7.

While the estimation strategy is not a replacement for other analyses, the estimator in this paper could be advantageous in certain circumstances. Such scenarios include whenever the data provide insufficient information on returnees' wages, such as in Census data (as in this paper and in Lacuesta, 2010; Ambrosini et al., 2015), when the sample size is too small to guarantee sufficient statistical power to the analysis (Ambrosini and Peri, 2012), or as a robustness check for the presence of feedback effects on the migration decision or Ashenfelter dip whenever pre-migration earnings are indeed observable, as in Fernandez-Huertas Moraga (2011), Ambrosini and Peri (2012), McKenzie and Rapoport (2010), and Kaestner and Malamud (2014).

5 Results

In addition to interest in the counterfactual estimation, the data enable studying different components of the return choice, as well as the wage determination process for Mexican-born immigrants in the U.S. Subsection 5.1 studies these choices, while Subsection 5.2 presents the density estimation results.

5.1 Parameter Estimates

The estimates of the marginal effects for the observable characteristics determining the decision to stay in the U.S. are presented in Table 1. Because these marginal effects are computed at the mean, the first column of the table reports the average characteristics of the immigrant sample.

[TABLE 1 HERE]

Each additional year of age has only a small effect on the probability of staying, increasing it by 0.4%. Compared with uneducated individuals, Mexicans who have completed primary (secondary) school are approximately 2% (6.7%) more likely to stay, while Mexicans with a college degree are approximately 3% more likely to stay. Having a foreign-born spouse slightly reduces the probability of staying, while individuals with a U.S.-born spouse are approximately 5% more likely to stay compared with those that have a foreign spouse. Meanwhile, having a foreign (U.S.)-born child reduces (increases) the probability of staying by approximately 6% (17.5%). It should also be noted that the two variables indicating social attachment to the host country are strongly significant, and – based on observable characteristics – stayers are more likely to have better educational outcomes. In addition, Figure 2 shows the actual and predicted share of stayers as we change the observable characteristics (as represented by decile of the index) and the distribution of the

predicted probability of staying in the U.S. Overall, the model fits the data well and the probability of staying in the U.S. takes a wide range of values.

Following Robinson's (1988) estimation, the procedure explained above also produces results for the wage process. These results are presented in Table C.2 in the Appendix and show standard labor market premia of the various individual characteristics. In estimating the counterfactual density of interest, I use a parsimonious specification where wages are estimated conditional only on demographics and socioeconomic characteristics such as educational attainment and family status. The results, based on a full set of controls, are reported as a robustness check in Section 7.

The estimation of the selection equation allows selecting those individuals for which sample selection disappears, namely, those individuals in the "high probability set". As mentioned in the previous section, individuals with a high probability of staying need to be selected. Table 2 shows the characteristics of the individuals inside and outside this high probability set. These individuals are on average more educated and more likely to have a U.S. born child and spouse. It should be noted that these differences will be taken into account when constructing the wage distribution in the next section. Instead, the main identifying assumption is that the selection of this subsample is exogenous. In the robustness checks section, I will further comment on the evidence of Z being exogenous.

5.2 Density Estimates

The following three research questions will now be answered in turn: (i) how different is the full immigrant population in terms of observable and unobservable traits compared with the population that stays in the U.S.; (ii) what would the distribution of wages be in the absence of return migration; (iii) how does this distribution change, conditional on educational characteristics?

How different is the immigrant population from the population of stayers in the U.S., in terms of observable and unobservable traits? Table 3 reports the deciles of the predicted wage, the residuals and the wage process that are observed and that would have been observed had there been no return migration. These quantities were calculated in the following manner. The first panel shows the predicted actual and counterfactual wages, both calculated as the product of the returns on the skills reported in Table C.2 and the characteristics of immigrant stayers' (immigrant population) characteristics for the actual (counterfactual) predicted wage distribution, i.e., $\hat{c} + \hat{\beta} X_j$, where j = only stayers, immigrant full population. I then report the average $\hat{c} + \hat{\beta} X_j$ in each decile of its distribution. The second panel reports the average estimated residuals in the actual and counterfactual distribution, by decile. As explained in Section 4.1, these were estimated from the high probability set. The last panel shows actual and counterfactual wages. Here I consider the distribution of the residuals and shifts it by the average predicted wages in the actual and counterfactual scenarios (that is, by the mean of $\hat{c} + \hat{\beta} X_j$). The deciles of the predicted wage are therefore reported as a summary measure. Because at all deciles the full population has higher unobservable characteristics than the selected population, as well as only marginally

lower observable skills, the wage distribution of the population will lie to the right of the observed distribution among the stayers, even if other quantiles were used.

[TABLE 3 HERE]

In terms of observable characteristics, Mexican immigrants would on average earn less had there been no return migration. In fact, the log-difference across the different deciles is always negative, which is in line with the descriptive analysis that found returnees to have below average skills. However, these differences are relatively small, reaching at most a decrease of a few cents (approximately 0.8%) in the wages between the two scenarios, because returnees represent only a small proportion of the total immigrant population.

The role of unobservable traits is shown in the second panel of Table 3. Unobservables were calculated as the difference between the actual and the predicted wages for the stayers, and were directly estimated for the full population using the estimation technique described in Section 4. I find that the positive differences between the counterfactual and actual distributions are driven by dissimilarities in unobservable traits. Had there be no return migration, the immigrant population would have been earning approximately 7.7% more (approximately 1 dollar, at the median) due to unobservable differences between stayers and returnees. The effect at the average level is a 4.5% change in wages, which is consistent with the relatively small effect of selection at the mean reported in previous studies (Lindstrom and Massey, 1994; Ambrosini and Peri, 2012).

The evidence presented suggests that immigrant stayers and the full population (i.e. stayers and returnees) are somewhat close in terms of observable traits, whereas differences arise in terms of unobservable traits. In particular, despite returnees being a disadvantaged group in the labor market in terms of observable traits, their unobservable abilities seem to compensate for this lack of skills. Furthermore, it seems that unobservable motives might push returnees to be more successful in the host country than the immigrants who stay. Although we cannot directly explain the motives behind returns, it is possible to conjecture that these immigrants leave the host country upon reaching their savings or skills acquisition goals and that more motivated immigrants are able to meet their personal objectives despite their original disadvantages in the host country labor market.[17]

What would the wage distribution be in the absence of return migration? The overall impact of return migration is presented in the last panel of Table 3, which reports the deciles of the actual wage distribution for stayers and those of the counterfactual wage distribution that would have occurred in the absence of return migration. In practice, this second distribution sums the average observable traits (panel one) and the unobservable components (panel two) for the immigrant population at each decile. As differences primarily arise due to the positive difference in the unobservable characteristics of the two populations, the implied counterfactual distribution

[17]Yang (2006) explores the reasons behind the returns of Filipino migrants and finds that while lifecycle considerations often motivate return migration, some migrants are motivated by target earnings considerations.

suggests that Mexican immigrants would be earning more had there been no return migration. In particular, more people would be earning above the median level.

Figure 3(a) presents the actual and counterfactual distributions described graphically in order to better visualize them. Although relatively close to each other, some differences in the two distributions are apparent from this figure. In the absence of return migration, more Mexican immigrants would appear in the upper tail of the distribution, thereby increasing the average wage in this population. To better observe this point, Figure 3(b) presents the difference between the counterfactual and actual distributions. Without return migration, more mass would appear in the upper tail of the wage distribution, as the wage difference is shown to be first negative and subsequently positive. Therefore, the disadvantage that returnees face in terms of lost human capital skills is balanced by the higher unobserved motivation and productivity displayed by this group. Overall, this balance translates into an increase in the concentration of individuals in the middle to upper part of the wage distribution in the absence of return migration. A Kolmogorov-Smirnov (KS) test for the difference between these two distributions delivers a D statistic of 3.05, implying that the actual and counterfactual distributions are different at all conventional significance levels.

[FIGURE 3 HERE]

This finding is not the only insight from the analysis, however. The last panel in Table 3 shows that return migration also affects wage inequality, reporting the 90-10, 90-50, and 50-10 wage gaps for the actual and counterfactual distributions. At the bottom of the distribution, the absence of return migration would imply an 7.3% increase in the difference between the 50th and 10th percentiles, whereas a reduction in this dispersion would occur at the top of the distribution. Overall, inequality within the Mexican population would increase slightly in the absence of return migration. Therefore, because selective return migration encourages high-wage earners to leave, this leads to a reduction in inequality within the Mexican population remaining in the U.S. By contrast, if all returnees were to stay, the full wage distribution in the population would display a slightly higher dispersion compared with that previously observed.

How does the wage distribution change conditional on educational characteristics? Since an individual's educational attainment greatly affects both his or her decision to stay and his or her wage, the importance of selection might vary by educational level. I therefore generate a distribution of unobservables specific to each education group. In other words, the unobservables are now based on a sub-sample of the high probability set with varying levels of education (primary, secondary, tertiary) and I compare these distributions to that of the actual stayers.[18] Accordingly, Table 4 reports the deciles of the predicted wage, unobservables and actual wage distributions for people with a primary school education, high school education and college degree. As before,

[18]These distributions are based on 448 observations for individuals with primary education, 6043 observations for the category of secondary educated and 180 observations for individuals with tertiary education. Because the vast majority of observations falls in the second category, it is unsurprising that the actual and counterfactual distributions are closer for this group.

the differences in observables are negligible across all educational groups, while unobservables are shown to drive dissimilarities in the wage process.

[TABLE 4 HERE]

However, despite the fact that returnees with primary- and secondary-level educations tend to show higher unobservable traits, the distribution of unobservables is different for college graduates. Figure 4 shows the dissimilarities in the actual and counterfactual distributions at different educational levels in order to better visualize these differences. Figures 4(a) and 4(b) first show the distribution of log-wages for low-educated individuals: as before, returnees are disproportionately drawn from the upper tail of the density. The same conclusion can be inferred from Figures 4(c) and 4(d), which show the same distribution for workers that have a secondary-level education. Finally, Figures 4(e) and 4(f) show what would have happened if all returnees with a college degree had stayed. In this case, a much larger mass of individuals would appear at the center of the distribution.

[FIGURE 4 HERE]

The findings presented above suggest two main conclusions. First, not all returnees are low-wage earners. In other words, within each educational group some returnees are high earners. Second, most of the action happens at the tails of the distribution: while almost no differences can be detected for individuals educated to secondary level, selective return migration has a much larger impact on individuals with either a low or high level of education.

6 Discussion and Policy Implications

In the absence of return migration, more Mexican immigrants would appear in the upper tail of the wage distribution. The results presented in Section 5 suggest that those immigrants who decide to leave are high-wage earners. Consequently, without return migration, the average wage in the population would be higher. This is not only true overall, but also when considering different education levels within the immigrant population. Despite the returnees being less skilled in comparison with stayers, they have higher unobservable traits that make them more successful in the labor market. This finding implies that an analysis that simply controls for differences in observable characteristics might draw the misleading conclusion that returnees are those who fail in the host country. On the contrary, returnees are not failures, but rather those who reached their goals in the host country, either in terms of savings or in terms of skills acquisition.

These important results extend the findings of Lubotsky (2007) and Hu (2000). In particular, Lubotsky (2007) shows that negative selectivity is less predominant in the Hispanic population, but the author was unable to explain this finding because of the impossibility of identifying the subsample of Mexican workers in the data. Furthermore, the results are also in line with the conclusions of Ambrosini and Peri (2012), who found indicative evidence of positive selection based on the pre-migration earnings of returnees compared with immigrant stayers despite the use of a

small sample. Finally, given that recent evidence on the selection of Mexican immigrants to the U.S. hints at negative selectivity (Ambrosini and Peri, 2012; Fernandez-Huertas Moraga, 2011), the results are in line with Borjas and Bratsberg's (1996) model in which selection on return migration intensifies the original selection process.

The policy implications of these findings are twofold. First, the assimilation process of Mexican migrants might have been underestimated due to selective out-migration. Second, if migration policies or economic conditions were to increase the length of stay, or even induce temporary migrants to settle permanently in the U.S., the consequences would not necessarily be of increased competition for immigrant and native low-wage earners.

Return migration influences immigrant inequality. The analysis presented has shown that return migration decreases inequality at the bottom of the distribution and increases inequality at the top. Therefore, the 90-10 wage differential changes only slightly. These effects are similar even when only taking account of individuals that have primary and secondary levels of education. The conclusion about high-skilled workers is different, however: return migration undoubtedly increases wage inequality within this group. Therefore, if policymakers are concerned about low earners, selective return migration seems to alleviate the dissimilarities in this population. However, if the goal of immigration reforms were to increase the average skills level of the incoming alien population, it should be recognized that the top earners of this group also return to their home countries.

Leaving aside general equilibrium effects, the immigrant–native wage gap would slightly close in the absence of return migration. An implication of this paper can be drawn by comparing the counterfactual distribution of wages with the wage distributions of native-born workers (the latter are shown in all the figures presented earlier).[19] Figure 3 shows that in the absence of return migration the immigrant wage distribution would become closer to the native-born wage distribution. The most interesting comparison can be observed in Figure 4, however, where the wage distribution is represented by educational level, demonstrating that all levels of human capital present a consistent earning gap between Mexican-born and native-born workers. This gap would close slightly for both very low and very high levels of education if all immigrants were to stay.

The difference between the actual, counterfactual and native-born wage distributions is also striking for high- and low-educated individuals for two reasons. First, Figure 4 clearly shows that selection on return migration is inducing middle to top earners to leave the U.S., thus biasing our picture of Mexican performance at both low and high levels of education. For example, in the absence of return migration, more of the top earners among low-skilled workers would stay in the U.S. A similar conclusion also holds for high-skilled workers. Therefore, a randomly selected Mexican immigrant would actually be doing better than shown herein. As an example, consider a migration policy that guarantees entry to the U.S. to individuals with high levels of education. This

[19] All the regression results and table for the native workers based on which the distributions are derived are available upon request.

policy still might not fully benefit the U.S. as middle to top wage earners - the most productive workers - would still leave.[20]

7 Robustness Checks

While Section 5 presented the main results of the paper based on the estimation of a parsimonious wage equation, this section checks the robustness of these results based on different model specifications. In particular, it controls for a fully specified model, estimates the model parametrically, discusses the effects of illegal and circular migration on the estimates, and discusses the validity of the assumptions of the estimator.[21]

[FIGURE 5 HERE]

Full Model. The previous discussion constructed the counterfactual and actual distributions based on the estimation of a parsimonious wage equation, only reporting those variables for which information was provided for both stayers and returnees.[22] There may be some concern, however, that a better specified model could change the results. Panel (a) in Figure 5 shows the actual and counterfactual distribution when more regressors are added into the analysis. Specifically, the wage process now includes controls for length of stay, industry and region of residence. Hence, the residuals based on which the actual and counterfactual densities are estimated should now exclude the effects of these variables. These variables were not included in the estimation of the selection equation as information on the migration experience is not available in the Mexican sample.[23] All previously drawn conclusions hold for this further specification, where the wage equation is better specified.[24]

Parametric Model. A fully semiparametric specification was adopted throughout the analysis in order to avoid inconsistency if the normality assumption was violated in the data. The same technique adopted for recovery of the population distribution of the error term u^*, however, can also be applied in a parametric setup.

Figure 5(b) show the actual and counterfactual distributions obtained by estimating a probit model to select for the individuals in a high probability set and the Heckman correction model is used to estimate the conditional expectations of the wage process.[25] The parametric results can

[20] I am implicitly assuming that this policy would not change the selection process of immigrants with high levels of education from Mexico to the U.S.

[21] The tables for the actual and counterfactual distribution are available upon request.

[22] These are shown in column 1 of Table C.2.

[23] The corresponding regression results are shown in column 2 of Table C.2

[24] As previously explained, the main problem with using a fully specified wage equation is the absence of information for returnees on the length of stay, the location, and the industry in the U.S. In the figure, I have assumed that returnees present similar characteristics to those of average non-returning migrants. Given the previous similarities of the deciles of $X\hat{\beta}$, this assumption seems reasonable.

[25] These results are shown in Table C.3, which presents the estimates for the decision to stay and the wage equation when both models have been estimated parametrically. A probit model was adopted to estimate the return choice.

be seen to be very close to the semiparametric results. Further, a KS test of the equality of the parametric and semiparametric models' distributions delivered a statistic of 1.13, which is below the 10% critical value, implying that the null of equality of distribution functions could not be rejected. This is not surprising because the wage process follows a log-normal distribution. This result is also reassuring as it shows that the technique applied can be easily implemented in a parametric setup. The parametric model has the further advantage of being efficient under normality, as shown by the smaller standard errors. However, it is important to note that the major conclusions of the paper are also valid in this context. Mexican returnees come from the middle to top part of the distribution, suggesting that a larger mass of people would have their wages in the upper part of the wage distribution in the absence of return migration.

Recent immigration. A key challenge in the dataset is the sample selection problem due to the choice of the two censuses. While we would ideally like to compare all stayers in the U.S. (captured in the U.S. census) with all returnees in Mexico, as mentioned in the Introduction and Data sections, only those individuals who returned between 1995 and 2000 are captured in the Mexican census. Hence, as return happens predominantly within 10-15 years since migration, the results might be driven by the comparison of recent return migrants with older immigration cohorts. To understand whether this limitation drives the conclusion, Figure 5(c) estimate the model on the subset of stayers that immigrated to the U.S. between 1990 and 2000, and as can be seen, such a restriction does not change the main conclusions.

Education. As previously mentioned, one of the drawbacks of combining the two censuses is the misreporting of education in the U.S. census. Mexican migrants in the U.S. might overstate their levels of educational attainment (Ibarraran and Lubotsky, 2007). If this were the case, any observed differences in educational attainment might be partly due to misreporting in the U.S. census rather than the selection of returnees. For this reason, educational attainment in the paper is measured in four broad categories (less than primary, at least primary, at least secondary education and college education) that do not distinguish between lower and upper secondary degrees. Moreover, the pattern shown in the data used herein can also be found in other studies that do not combine these two censuses (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012), which could reduce such concerns. To further check whether this is a serious challenge to the results, the range of the educational variable was further reduced to two categories (college and non-college graduates). The corresponding actual and counterfactual distribution are reported in Figure 5(d), and once again, results are qualitatively unchanged.

The first column of the table reports the implied marginal effects, which are very close to the semiparametric marginal effects. The second column of Table C.3 shows the results for the wage equation. Following the same logic used for the semi-parametric estimator, I have then constructed $\hat{u}^*$ as the vector of residuals for individuals in the top 95th-percentile of the probability of staying, now defined by the cumulative normal distribution evaluated at the index in the S_i decision. I finally compared the distribution of wages implied by this sample, where selection had been removed, with the distribution of wages in the selected sample.

Undercount of migrants in the U.S. Census. The U.S. census is known to undercount immigrants, especially if they are undocumented and short-term stayers (Hanson, 2006; Fernandez-Huertas Moraga, 2011). The fear of deportation or the fact that migrants are not physically present in the country at the time of the enumeration might induce this population not to complete the census form. Consequently, the census sample might not represent the actual Mexican population in the U.S. For instance, Fernandez-Huertas Moraga (2011) shows that the results of the positive selection in Chiquiar and Hanson (2005) are largely driven by the non-representativeness of the Mexican sample in the U.S. census due to the two causes mentioned above.

These problems might be particularly severe when using U.S. census data. By contrast, it seems reasonable to assume that Mexican returnees are better captured in the Mexican census as the motivation to underreport U.S. experience is not affected by the illegality status in the U.S. (although the presence of short trips might still induce an undercount of migrants in this sample).

The resulting key concern is the non-randomness of the census sample even after controlling for relevant observable characteristics. In the following discussion, I argue that under certain conditions the identification strategy adopted in the analysis is robust to the non-randomness of the U.S.-Mexican sample.

To visualize the effects of an undercount of Mexican immigrants in the U.S. census, let $C_i = 1$ be an indicator that equals 1 if the respondent appears in the census and 0 otherwise:

$$C_i = \begin{cases} 1 & W_i'\gamma \geq \eta_i \\ 0 & W_i'\gamma < \eta_i. \end{cases}$$

Subsequently, the choice of staying in the U.S. is only observed if the individual appears in the census:

$$S_i = \mathbf{1}(Z_i\alpha \geq \epsilon_i) \qquad \text{if} \qquad C_i = 1.$$

The concern is that η_i and ϵ_i are correlated and, in particular, based on the results presented by Passel (2006), we expect them to be positively correlated: individuals who are more likely to appear in the sample are also those more likely to stay and to have longer U.S. experience. If η_i and ϵ_i are correlated, then there might be a concern that the probability of staying $P(S_i = 1|Z_i'\alpha)$ has been mis-estimated. Using again the Law of Total Probability, in fact:

$$\begin{aligned} P(S_i = 1|Z_i'\alpha) = P(S_i = 1|Z_i'\alpha, C_i = 1)Pr(C_i = 1|Z_i'\alpha) + \\ + P(S_i = 1|Z_i'\alpha, C_i = 0)Pr(C_i = 0|Z_i'\alpha), \end{aligned}$$

where the second part of the addition is missing. However, note that the high probability set was constructed by sending $P(S_i = 1|Z_i'\alpha, C_i = 1)$ to 1. By doing so, individuals with large values of $Z_i'\alpha$ were implicitly selected. However, high values of $Z_i'\alpha$ are associated with high values of $W_i'\gamma$. In fact, the main variable that can send that probability to 1 is age. For instance, older individuals are not only more likely to stay but also more likely to be captured by the census (Passel, 2006). Thus,

whenever $P(S_i = 1|Z_i'\alpha, C_i = 1)$ is close to 1, $Pr(C_i = 1|Z_i'\alpha)$ is also expected to be close to 1. This implies that in the high probability set the probability of staying is largely determined by individuals who do appear in the sample, i.e., $P(S_i = 1|Z_i'\alpha) \approx P(S_i = 1|Z_i'\alpha, C_i = 1)Pr(C_i = 1|Z_i'\alpha)$. As a consequence of using individuals in the high probability set, the distribution of the unobservables recovered should be unaffected by illegal immigration. In other words, given the relation between S_i and C_i in this particular application, using the high probability set appears to marginalize the problems related to the censoring in the selection rule due to illegal immigration.

As a final check, focusing on concerns of illegal migration, as most illegal migrants seem to settle in California, Florida, Texas and New York (U.S. Immigration and Naturalization Services, 2000), I run the analysis excluding these states and observing how conclusions change. The actual and counterfactual wage distributions are reported in Figures 5(e): dropping states where illegal migration may be predominant does not change the main conclusions of the paper. Further, the difference in the two distributions becomes more marked, as expected in the case of an undercount of migrants in the U.S. that have worse labor market outcomes. As a final remark, several analyses report a much lower undercount in the 2000 census (Card and Lewis, 2007) than in previous enumerations.

Circular Migration. Mexican migrants are recognized for engaging in repeated movements into the U.S. (Massey and Espinosa, 1997). Using the Mexican Migration Project, and constructing a sample similar to that used in the analysis, it seems that circularity is more predominant in communities residing at the U.S.-Mexico border.[26] Hence, the model was run excluding the bordering states (panel (f) of Figure 5). As shown, the results are unchanged in this specification.

8 Additional Checks on the Identifying Assumptions

The estimation strategy employed in this paper is based on a set of assumptions. This section discusses whether they are likely to hold. First, it discusses the results by showing that using different sources and different techniques findings are still compatible with those reported in the data. Then, it discusses the challenges of sampling returning migrants. Lastly, it directly assesses the validity of the identifying assumptions of the empirical strategy in the particular context presented in this paper.

Additional Evidence on the Validity of the Assumptions using the Mexican Family Life Survey. To address whether the results are driven by the adopted estimation strategy and its underlying assumptions, I provide additional evidence on the selection patterns of returning migrants based on the Mexican Family Life Survey (MxFLS). Although these additional findings are based on a very small sample size, a key reason why return migration poses serious challenges

[26] Over half (53%) of circular migrants reside in a border state, compared with 43% for the single-trip migrants. Hence, individuals that reside outside bordering states are significantly less likely to be circular migrants at the 5% level. These results available upon request.

when analyzed, the similarity of the results obtained gives reassurance about the validity of the technique adopted in this paper.

The MxFLS is a longitudinal dataset which collects information on socioeconomic indicators, demographics and health indicators of about 8,000 Mexican households. The baseline (MxFLS-1) was conducted during 2002. The second wave of field work (MxFLS-2) was conducted during 2005-2006 with a 90 per cent re-contacting rate at household levels. This dataset contains in principle the ideal structure to study the selection in return migration as migrant and returnees are followed across borders. In practice, sample sizes are too small (about 130 cases) to make it the key source for studying return migration. Ambrosini and Peri (2012) have already highlighted this problem, however still commenting on key selection patterns in this group. Following their work, I compare in Figure 8(a) income levels of U.S. stayers and Mexican returnees *before* migration. The key advantage of this strategy is twofold. First, it follows the most recent approach to measuring selection by basing it on pre-migration outcomes (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012; Kaestner and Malamud, 2014).[27] Income here is intended as a reduced form representation of a worker's productivity, including both observable and unobservable factors. The focus on pre-migration income allows therefore obtaining a summary measure of the economic potential of migrants, independently of the additional challenges that will be faced in the host country. Second, the distribution of pre-migration income is a direct, non-parametric measure of selection, that avoids the complex exercise undertaken in the paper and does not rely on the assumptions that are now under scrutiny.

[FIGURE 8 HERE]

Using the MxFLS I have identified individuals who migrate between 2002 and 2005 and, among them, identified those who are still in the US in 2005 (U.S. Stayers, $S = 1$) and those who have returned to Mexico by 2005 (returnees, $S = 0$). There are 122 individuals who migrated to the U.S. between 2002 and 2005, of which 27 had already returned to Mexico by the time of the interview. Figure 8(a) shows pre-migration income of both U.S. stayers and of the full population of migrants. The figure shows that returnees are positively selected based also on pre-migration income. The comparison of stayers' income with the full population provides similar results to those obtained from the census. In other words, U.S. stayers are negatively selected compared to the full migrant population and this is true even before migration, i.e. before the migrants had been exposed to U.S. shocks. This check is indicative for two reasons. First, it should convince that the main results are not driven by the identifying assumptions of the paper. Second, it also suggests that returning migrants are not (only) a sample of individuals with failed U.S. experience as observational differences in terms of productivity are present even before migration. In particular, looking at the probability mass above the mean, data from the MxFLS suggests that the difference in the two distributions peaks at 5%, while in the main dataset the difference peaks at about 10%. Returning

[27] I rely on individual income due to missing values in wages, which would further reduce the already limited sample size. Results do not change if wages are used instead of income.

migrants are doing better than pre-migration income would suggest. This conclusion seems in line with return migration being part of a life-cycle choice.

Sampling Return Migrants. While sampling of the Census is unlikely to be related to the decision to stay or return, the strategy of keeping migrants/stayers only and then pooling the two datasets might raise concerns of choice-based sampling. Although this is commonly done in the literature (Borjas and Bratsberg, 1996; Chiquiar and Hanson, 2005; Lacuesta, 2010; Mishra, 2007; Caponi, 2011), one might be concerned that the returning probabilities and returnees' characteristics obtained in this pooled sample might differ from the characteristics and probabilities in the population of migrants at risk of return.

The results based on the MxFLS dataset already indicated that similar patterns could be obtained in a sample not subject to this particular challenge.

To gain further knowledge about the severity of this concern, I have explored additional information provided in a supplementary migration module of the Mexican census. The head of the household was asked to give information on the number of household members who had left Mexico between 1995 and 2000 as well as the number of household members who had left in that same time period and returned by 2000. Unfortunately, no information was collected about the experiences and characteristics of these stayers and returnees and, additionally, it is not possible to accurately merge individual information from the full census with information from the migration supplement because unique individual-level identifiers are not provided. It is therefore impossible to know how the (observed and unobserved) characteristics of stayers and returnees would compare with those of the pooled sample used in this paper. Lastly, this data will not record individuals whose complete household emigrates to the U.S. and no member returns to Mexico, estimates of which are about 8% (Fernandez-Huertas Moraga, 2011). Nonetheless, this source can be used to calculate return rates of individuals who have left between 1995 and 2000.

This source suggests that 25.96% of migrants had returned in that five-year period. Correcting for whole migration household, the return rate from this source comes close to the 17.18% return rate estimated by pooling the Mexican and the U.S. Censuses for that particular five-year interval (see also Figure 5(c)).

Lastly, while individual-level analysis is not possible, it is still possible to analyze average income in Mexico for households with all migrants members currently in the U.S. and households with returning members. Figure 8(c) compares the different levels of income for all households (with all migrant members still abroad and with at least a returning migrant) and for households with U.S. stayers only. Albeit smaller, as we would probably expect from a household-level analysis, selection patterns are in line with those in the paper.

Validity of the Assumptions in the Estimation. Before concluding, it is important to discuss the several identifying assumptions based on which the model is estimated. I start by discussing the validity of having a U.S.born child as an exclusion restriction and then continue by discussing endogeneity of the regressors and heteroskedasticity in this particular application.

The empirical analysis uses having a U.S.born child as the variable entering the selection equation and excluded from the outcome equation. The reader might be concerned about the sensitivity of the procedure to the use of this particular exclusion restriction. As a first step, it should be noted that the literature on selection models has recently advanced to develop tests of the key identifying assumptions in sample selection models (e.g. Blundell et al., 2007; Kitagawa, 2010; Huber and Mellace, 2014). I use the test developed in Huber and Mellace (2014) which assesses the joint satisfaction of the validity of the exclusion and of (the usually less discussed) additive separability of the error term in the selection equation. The two assumptions of sample selection models imply two testable inequality constraints that come from both point identifying and bounding the outcome distribution of the subpopulation that is always selected. Following the procedure suggested by these authors, I test the resulting constraints on the mean outcome of the always selected. The null hypothesis is:

$$H_0 : \begin{pmatrix} E(Y|USborn = 1, S = 1, Y < Y_q) - E(Y|USborn = 0, S = 1) \\ E(Y|USborn = 0, S = 1) - E(Y|USborn = 1, S = 1, Y \geq Y_{1-q}) \end{pmatrix}$$
$$\equiv \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where q corresponds to the proportion of always selected in the mixed population of individuals with and without US-born children (83% in my sample) and Y_q represents the qth conditional quantile in the conditional outcome distribution given $USborn = 1$ and $S = 1$. The standardized mean constraints are (-0.1073, -0.2434). Such negative value indicates that the inequalities are never binding. Hence, we do not have enough evidence pointing to a violation of the identifying assumptions in a sample selection model when relying on having a US-born child as an instrument. [28]

To further extend the discussion on the exclusion restriction, it is worth noting two additional issues. First, the marginal effects from the baseline model (Table 1) are close to the marginal effects of a probit model (Table C.3), suggesting that the errors in the selection and outcome equations might be jointly normally distributed. Normality is theoretically sufficient for identification. Relying only on the non-linearities of the selection model, Figure 6(a) shows the counterfactual obtained in a parametric framework that does not use having a U.S.-born child as an exclusion restriction. Results are consistent with the benchmark analysis.

[FIGURE 6 HERE]

Second, technically, identification at infinity does not require the use of an exclusion restriction (Chamberlain, 1986). Intuitively, as identification at infinity correctly identifies a subset of the data in which sample selection does not appear, all parameters of the model could be estimated directly on that subset without resorting to more complex exercises. Identification is instead reached

[28] The associated p-value is 1, so we fail to reject the null of validity/monotonicity at conventional statistical levels.

through regressors having a larger probability mass at the tails compared with the error term (Andrews and Schafgans, 1998; Klein et al., 2015) and the assumptions discussed in Section 4.3 need to hold. Mulligan and Rubinstein (2008) follow Chamberlain's suggestion and estimate a wage equation directly in a subset of their data where females have a probability of participating in the labor force above the 95th percentile, without resorting to an exclusion restriction. Following this same strategy, complementary models that estimates the full set of parameters and the counterfactuals directly in the high-probability set were run, and these provided similar results to those presented in the paper (results available upon request).

Besides the exclusion restriction, the adopted technique assumes exogeneity and homoskedasticity. Exogeneity is needed for consistency. In fact, the assumed exogeneity of the regressors Z in model (1) from u^* guarantees the randomness of this selection rule. The exogeneity of X is naturally important for the correct identification of the parameters in the wage equation but such importance is not specific to the current estimation strategy, thus, in the following I will focus on the role of Z. It should be noted, however, that Z and X coincide with the exclusion of one variable, hence the following discussion and tests closely relate also to the assumption of exogeneity of X.

To better visualize the problem, let G be the control function that accounts for the bias that would occur under selection into the decision to stay, i.e. a non-parametric equivalent of the inverse Mills ratio. In the population, G is an unknown function of $Z_i'\alpha_0$. The wage equation in the selected sample can be written as:

$$Y_i = X_i'\beta_0 + c_0 + G(Z_i'\alpha_0) + u_i^* \equiv X_i'\beta_0 + c_0 + \eta_i, \quad \text{with} \quad G(Z_i'\alpha_0) + u_i^* \equiv \eta_i \tag{5}$$

If sample selection was not considered and the presence of $G(Z_i'\alpha_0)$ was ignored, not only would the conditional expectation of Y_i on X_i be incorrectly estimated, but all moments of η_i would depend on Z_i. Therefore, we would expect the distribution of the residuals in the selected sample to change at different quantiles of $Z_i'\alpha_0$. To underline this point, Figure 6(b) shows this distribution and highlights the variation in the residuals as a function of different quantiles of the selection index.

On the contrary, if the estimation technique was correctly able to purge out sample selection when focusing on the high probability group (hence, obtaining $G(Z_i'\alpha_0) \approx 0$), the distribution of η_i would be equivalent to the white noise error term in the population, u^*, and, hence, would not change when conditioning on different quantiles of $Z_i'\alpha_0$. To provide evidence of whether exogeneity of Z_i is appropriate in this context, I estimate $\hat{f}(\eta_i|Z_i'\hat{\alpha})$ in the high probability set, which should be close to the estimated $\hat{f}(u^*)$ in the paper if Z_i can be treated as exogenous.

Figure 6(c) shows how 'close' the estimated $\hat{f}(u^*)$ is to the distribution of the unobservables in the high probability set conditioning on $Z_i'\alpha_0$. Conditioning on different quantiles of the index $(Z_i'\hat{\alpha})$ does not induce a considerable change in the distribution of u^*. Specifically, it seems that the recovered distribution $f(u^*)$ is relatively conservative, as it shows higher variability compared with the conditional distributions.

To provide further indicative evidence of this finding, a KS test can be used to test whether

these distributions come from the same underlying density. Table 5 tabulates the D statistic at different deciles of the index. The null that the conditional and unconditional densities are drawn from the same distributions cannot be rejected at the 5% significance level in the central deciles.

[TABLE 5 HERE]

Since the null is rejected at some deciles, Figure 7 uses the distribution in the high probability set conditional on the ninth decile of the $(Z_i'\hat{\alpha})$ index as the counterfactual of interest, as the KS test found this to be the furthest from the estimated distribution reported in the paper. As can be seen, the results remain consistent with the middle to positive selection of returnees and, if anything, results are stronger. This comparison hints that selecting on $(Z_i'\hat{\alpha})$ does not overturn the conclusions.

[FIGURE 7 HERE]

It should be noted that, if the excluded variable did enter the wage equation and thus was an invalid exclusion restriction, the estimated density for individuals in the high probability set would also change at different quantiles of the estimated index $Z_i\hat{\alpha}$. In fact, an invalid would cause spurious correlations between the error term in the wage equation and the observable characteristics, and such correlations would still be present in the high probability set. The results just shown should further corroborate the validity of the exclusion restriction.

Lastly, turning to the assumption of homoskedasticity, the variance structure of the model can be extended to allow for an unknown form of heteroskedasticity, at some cost to analytical tractability. This extension seems to be important, however, as conditional heteroskedasticity is common in empirical applications. The Appendix extends the estimation technique to allow for dependence in the second moment between the error and regressors. The results are in line with those reported in the main part of the paper, implying that even allowing for heteroskedasticity does not change the main conclusions.

9 Conclusions

The political discussion generated by Mexican migration flows into the U.S. has focused on understanding migration decisions, yet has until recently ignored the role played by selective return migration in shaping estimates of immigrant labor market outcomes. Indeed, relatively few previous studies have examined the breakdown between returnees and stayers in the host country. This paper thus adds evidence on this topic by recovering a counterfactual wage distribution in the absence of return migration. The estimation procedure presented herein extended the estimator in Andrews and Schafgans (1998) to its density counterpart, showing the overall distribution of wages that would be observed if all migrants were permanent and if such a distribution were conditional on educational attainment.

The results suggest that selective return migration improves average earnings of immigrants and reduces immigrant wage dispersion. Further, return migration has a greater impact on the tail

of the wage distribution. In particular, more mass would appear in the upper part of the wage distribution in both very low and very high educational groups in the absence of return migration, implying up to a 7.7% increase in the median wages and a 4.5% increase in mean wages paid to the Mexican migrant population. These results are stable across different wage specifications, samples and techniques. The impact at the mean is, however, relatively small, which might be the reason for the inconclusive findings presented in the literature. Our notion of Mexican migration has been distorted by selective return migration. Further, the presented results contrast with the general perception that those migrants who return have failed in the host country and with the findings of previous studies concerning the nature of return migration in the U.S.

Tables

Table 1: Marginal effects of variables on the Probability of Staying in the U.S., Mexican Born Men, 25-55 Years old

	Average Characteristics	Marginal Effects
Age	35.977	0.004***
		(4.29E-04)
Primary	0.468	0.020***
		(0.002)
Secondary	0.324	0.067***
		(0.003)
College	0.043	0.031***
		(0.002)
Married	0.796	-0.006***
		(0.001)
US born spouse	0.105	0.050***
		(0.002)
Child	0.629	-0.063***
		(0.003)
US born child	0.568	0.175***
		(0.004)
N	133,389	133,389

Standard errors in parentheses.
The fraction of stayers is about 89% and the average predicted probability of staying is 0.90.
Significance levels: *: 10%, **: 5%, ***: 1%.
The marginal effects are calculated at the average X and for a unit change from 0 to 1 for dummy variables.

Table 2: Demographic and socio-economic characteristics, High Probability Set vs Others

Variable	$H_i = 1$	$H_i = 0$	**T-test**
Average Predicted Probability	0.999	0.896	6.154
	(0.037)	(0.305)	
Age	38.215	35.859	2.041
	(9.751)	(7.788)	
Less than Primary School	0.003	0.174	-8.076
	(0.059)	(0.379)	
Primary Education	0.067	0.489	-11.026
	(0.250)	(0.500)	
Secondary Education	0.903	0.293	15.005
	(0.296)	(0.455)	
College Education	0.027	0.044	-0.836
	(0.161)	(0.205)	
Married	0.826	0.794	0.661
	(0.379)	(0.404)	
US born spouse	0.656	0.075	10.691
	(0.475)	(0.264)	
Child	0.677	0.627	0.872
	(0.468)	(0.484)	
US born child	0.977	0.546	13.769
	(0.151)	(0.498)	
N	6692	126697	

H_i is an indicator for the probability of staying being above the 95th percentile. See text for further details.

Table 3: Deciles of $\hat{Y}_i$ and $\hat{u}_i$ and Y_i, Parsimonious Model, Mexican-Born Men 25-55 Years Old.

Decile	**Actual**	**Counterfactual**	**Difference**
		Observables	
1	2.288	2.277	-0.011
2	2.345	2.344	-0.001
3	2.384	2.382	-0.002
4	2.424	2.417	-0.007
5	2.457	2.454	-0.003
6	2.483	2.479	-0.004
7	2.512	2.508	-0.004
8	2.552	2.544	-0.008
9	2.618	2.617	-0.001
Average	2.451	2.451	-4.773E-04
		Unobservables	
1	-0.670	-0.666	0.004
2	-0.492	-0.461	0.031
3	-0.350	-0.292	0.058
4	-0.221	-0.147	0.075
5	-0.097	-0.020	0.077
6	0.031	0.106	0.076
7	0.170	0.249	0.079
8	0.345	0.418	0.073
9	0.615	0.657	0.042
Average	-0.045	1.878E-07	0.045

Continue to next page

Continued from previous page

Decile	**Actual**	**Counterfactual**	**Difference**
	Log-Wage		
1	1.781	1.785	0.003
2	1.959	1.990	0.031
3	2.101	2.159	0.058
4	2.230	2.304	0.074
5	2.354	2.431	0.077
6	2.482	2.557	0.075
7	2.622	2.700	0.079
8	2.796	2.869	0.073
9	3.067	3.108	0.041
Average	2.406	2.451	0.045
	Inequality Measures		
10-90 Wage	1.285	1.323	0.038
10-50 Wage	0.573	0.647	0.073
50-90 Wage	0.712	0.677	-0.036

The first column (*Actual*) shows $\hat{Y}$ and $\hat{u}$ for the observed sample of stayers. The second column (*Counterfactual*) shows $\hat{u}$ if all returnees had stayed, hence refers to these quantities estimated in the high probability set; it also shows the observable characteristics of the full sample, corresponding to the observables for both stayers and returnees.

The number of observations used to calculate the actual distribution is 120,205. The counterfactual distribution of unobservables is based on 6,692 observations. The counterfactual distribution of observables is based on the full set of characteristics of stayers and returnees, hence 133,389 observations.

Table 4: Deciles of $\hat{Y}_i$ and $\hat{u}_i$ and Y_i by Education Level, Parsimonious Model, Mexican-Born Men 25-55 Years Old.

Decile	**Act.**	**Counterfact.**	**Diff**	**Act.**	**Counterfact.**	**Diff**	**Act.**	**Counterfact.**	**Diff**
	Primary Education			Secondary Education			College Education		
					Observables				
1	2.249	2.241	-0.008	2.356	2.356	0.000	2.672	2.672	0.000
2	2.305	2.299	-0.006	2.406	2.406	0.000	2.720	2.720	0.000
3	2.355	2.352	-0.003	2.462	2.462	-0.001	2.767	2.767	0.000
4	2.387	2.383	-0.003	2.495	2.495	0.000	2.810	2.810	0.000
5	2.417	2.406	-0.011	2.524	2.524	0.000	2.850	2.850	0.000
6	2.445	2.436	-0.009	2.554	2.554	0.000	2.885	2.880	-0.005
7	2.469	2.469	0.000	2.580	2.580	0.000	2.906	2.906	0.000
8	2.499	2.495	-0.004	2.607	2.607	0.000	2.931	2.927	-0.004
9	2.519	2.517	-0.002	2.635	2.635	0.000	2.945	2.945	0.000
Average	2.401	2.397	-0.004	2.512	2.511	-0.001	2.829	2.828	-0.001
					Unobservables				
1	-0.652	-0.666	-0.014	-0.674	-0.663	0.011	-0.879	-0.857	0.022
2	-0.488	-0.489	-0.002	-0.474	-0.459	0.015	-0.630	-0.476	0.155
3	-0.356	-0.329	0.027	-0.316	-0.291	0.024	-0.419	-0.170	0.249
4	-0.234	-0.190	0.044	-0.183	-0.146	0.036	-0.241	0.040	0.282
5	-0.119	-0.064	0.054	-0.053	-0.021	0.033	-0.068	0.177	0.245
6	0.007	0.061	0.054	0.076	0.105	0.028	0.104	0.297	0.193
7	0.143	0.228	0.085	0.214	0.243	0.028	0.274	0.444	0.169
8	0.313	0.422	0.109	0.385	0.410	0.025	0.471	0.639	0.168
9	0.585	0.622	0.037	0.633	0.654	0.020	0.758	0.875	0.117
Average	-0.055	-0.034	0.021	-0.022	0.000	0.022	-0.053	0.105	0.158
					Log-Wage				
1	1.749	1.731	-0.018	1.838	1.849	0.010	1.950	1.971	0.021
2	1.914	1.908	-0.006	2.038	2.052	0.014	2.199	2.353	0.154
3	2.045	2.068	0.023	2.196	2.220	0.024	2.410	2.658	0.249
4	2.167	2.207	0.040	2.330	2.365	0.035	2.587	2.868	0.281
5	2.283	2.333	0.050	2.459	2.491	0.032	2.761	3.005	0.244
6	2.408	2.458	0.050	2.588	2.616	0.027	2.933	3.125	0.192
7	2.544	2.625	0.081	2.726	2.754	0.027	3.103	3.272	0.169
8	2.714	2.819	0.105	2.897	2.921	0.024	3.300	3.467	0.167
9	2.986	3.019	0.033	3.146	3.165	0.020	3.587	3.703	0.116
Average	2.346	2.363	0.017	2.490	2.511	0.021	2.776	2.934	0.157
					Inequality Measures				
10-90 Wage	1.237	1.288	0.052	1.307	1.316	0.009	1.637	1.732	0.095
10-50 Wage	0.533	0.602	0.069	0.620	0.642	0.022	0.811	1.034	0.223
50-90 Wage	0.703	0.686	-0.017	0.687	0.674	-0.012	0.826	0.698	-0.128
N	54,651	54,651		41,694	41,694		5,202	5,202	

Act. shows $\hat{Y}$ and $\hat{u}$ for the observed sample. *Counterfact.* shows $\hat{Y}$ and $\hat{u}$ if all returnees had stayed. Therefore, the observable characteristics of the sample correspond to the observables for both stayers and returnees. The unobservables correspond to the predicted u^*. The density for the actual distribution is obtained by conditioning the residuals from the wage distribution to people with primary, secondary and tertiary education. The density for the counterfactual distribution corresponds to the counterfactual distribution of Table 3.

Table 5: Kolmogorov-Smirnov Test for the difference in the estimated counterfactual distribution and the distribution of residuals in the high probability set conditioning on different deciles of $Z_i'\hat{\alpha}$.

	D Statistic
Decile 1	1.83^{***}
Decile 2	0.61
Decile 3	1.74^{***}
Decile 4	1.23^{*}
Decile 5	1.16
Decile 6	1.13
Decile 7	1.09
Decile 8	1.38^{***}
Decile 9	2.37^{***}
Decile 10	2.21^{***}

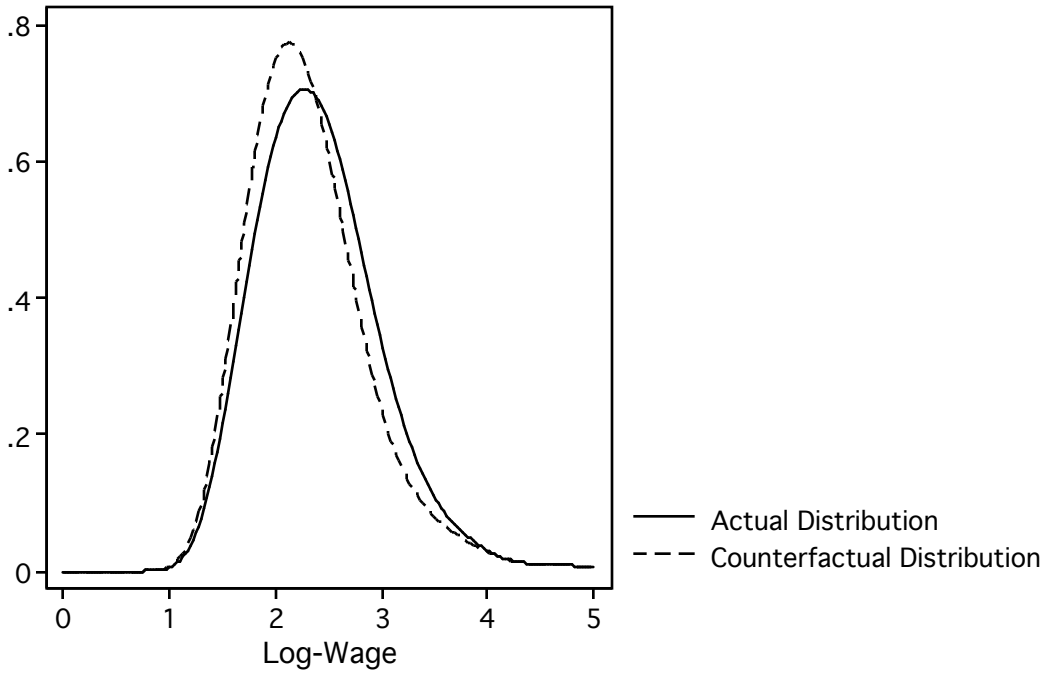
Significance levels: *: 10%, **: 5%, ***: 1%.

Critical Values: 10%: 1.22; 5%: 1.36; 1%: 1.63;

The test was constructed comparing the estimated distribution of u^* ($\hat{f}(u^*)$) with the distribution of the residuals for individuals in each decile of the $(Z_i'\alpha)$-index. See text for details.

Figures

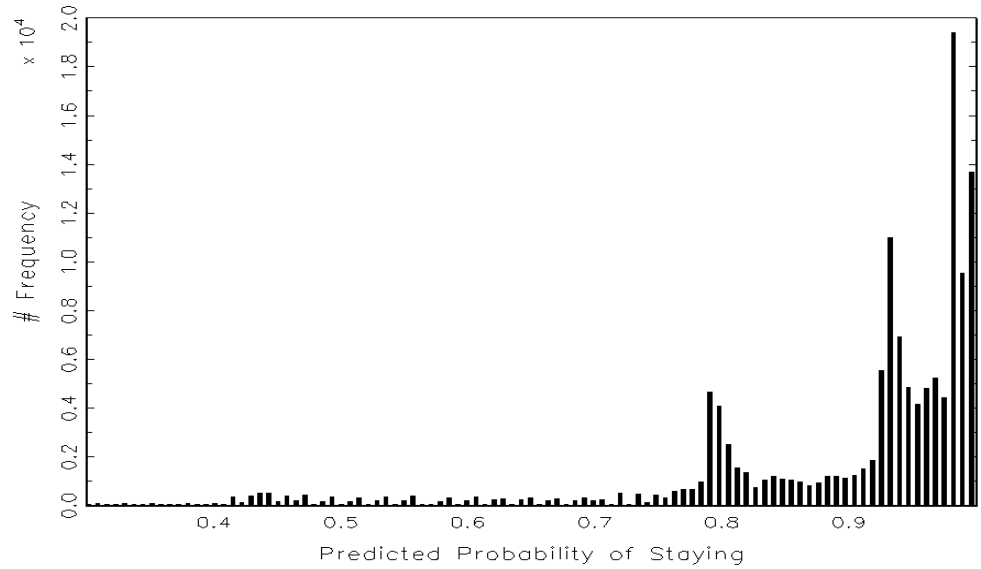
Figure 1: Wage Densities in the U.S., using DiNardo et al. (1996), Men, 25-55 Years Old.



The actual distribution represents the distribution of wages for Mexican-born workers in the U.S. The counterfactual distribution represents the distribution of wages that would have occurred in the U.S. if Mexican-born workers had the characteristics of the returnees. These distributions are obtained following DiNardo et al. (1996). To construct the counterfactual, a probit model is estimated on the probability of staying in the U.S. using the sample of Mexican-born workers in the U.S. and the returnees in Mexico. This model relates the probability of staying with age, dummy variables for schooling, marital status, having a child, and birthplace of the child and the spouse. The Gaussian kernel function with optimal bandwidth was used (Silverman, 1986), to be coherent with the reminder analysis of the paper.

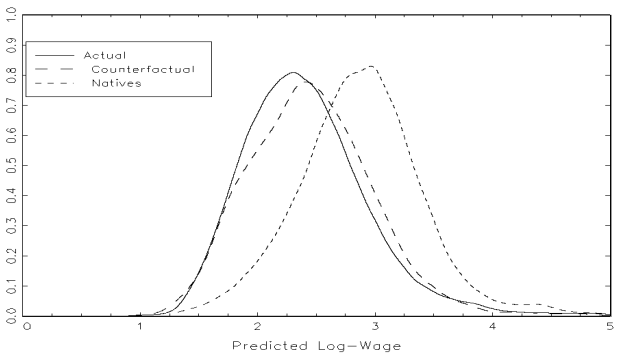
Figure 2: Actual and Predicted Probability of Staying in the U.S., Men, 25-55 Years Old.

(a) Predicted and Actual Share of Stayers, by index deciles

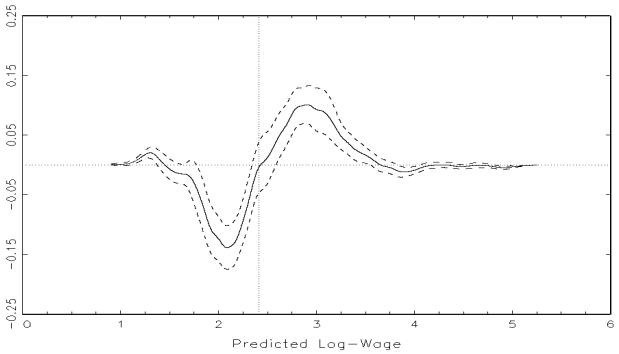


(b) Distribution of the Predicted Probability of Staying in the U.S.

Figure 3: Actual and Counterfactual Log-Wage Distributions (a) and their Difference (b), Parsimonious Model, Men, 35-55 Years Old.



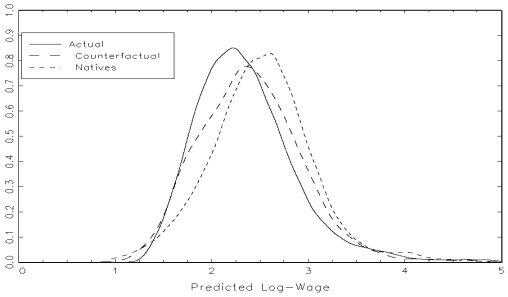
(a) Actual (Solid Line) and Counterfactual (Dashed Line) Wage Distribution



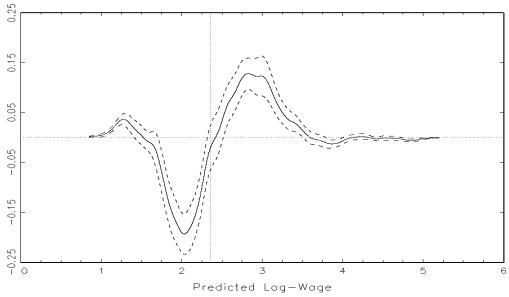
(b) Difference in Counterfactual and Actual Distributions

The Actual distribution represents the distribution of wages for Mexican-born workers currently residing in the U.S. The Counterfactual distribution represents the distribution of wages in the U.S. for Mexican-born workers if all migrants settled permanently, i.e., if no return migration occurred between 1995 and 2000. Section 4 in the paper explains how to recover these densities. Table 3 shows the deciles of these distributions. Standard errors have been bootstrapped (100 repetitions).
Kolmogorov-Smirnov test statistic for equality in the Actual and Counterfactual distribution: 4.95. Critical value at 1%: 1.63.

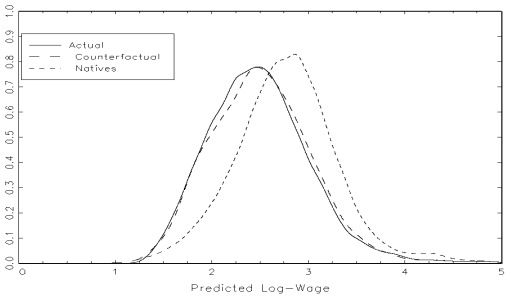
Figure 4: Estimated Actual and Counterfactual Log-Wage Densities for Mexican immigrants with Primary (a), Secondary (c), and College Education (e), Parsimonious Model, Men 35-55 Years Old.



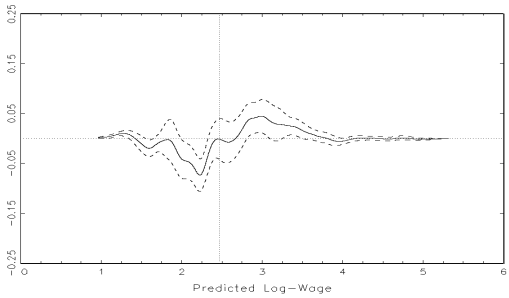
(a) Primary Education, $n = 27,403$



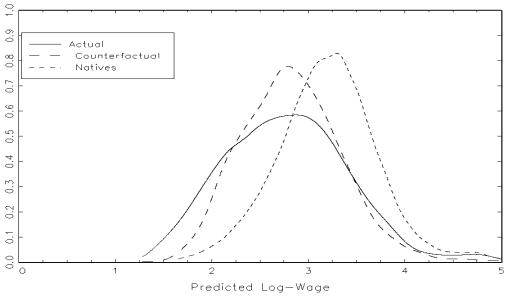
(b) Difference in Counterfactual and Actual Distributions, Primary Education



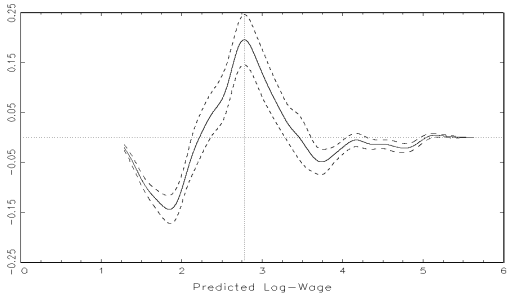
(c) Secondary Education, $n = 18,811$



(d) Difference in Counterfactual and Actual Distributions, Secondary Education



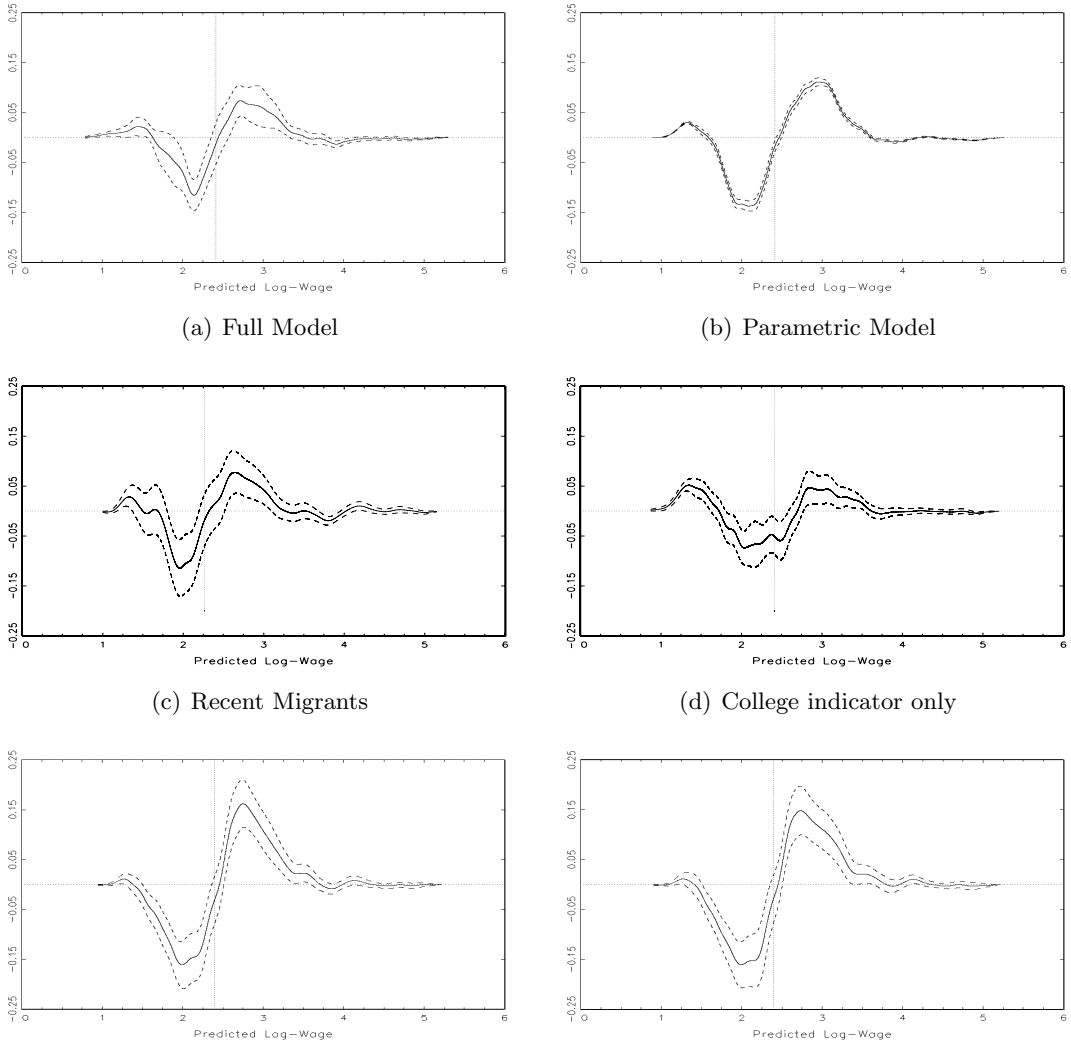
(e) College Education, $n = 3,001$



(f) Difference in Counterfactual and Actual Distributions, College Education

The Actual distribution represents the distribution of wages for Mexican-born workers currently residing in the U.S. and having primary (a), secondary (c) and tertiary (e) education. The Counterfactual distribution represents the distribution of wages in the U.S. for Mexican-born workers having primary (a), secondary (c) and tertiary (e) education if all migrants settled permanently, i.e., if no return migration occurred between 1995 and 2000. The density for the actual distribution is obtained by conditioning the residuals from the wage distribution to people with primary, secondary and tertiary education. The density for the counterfactual distribution corresponds to the counterfactual distribution of Figure 4(a). Section 4 in the paper explains how to recover these densities. Table 4 shows the deciles of these distributions. Standard errors have been bootstrapped (100 repetitions).
KS test statistic for equality in the Actual and Counterfactual distribution for Mexican-born workers with primary education: 6.16, with secondary education: 2.24, with tertiary education: 4.86. Critical value at 1%: 1.63.

Figure 5: Differences in Counterfactual and Actual Log-Wage Densities, Full Model (a), Parametric Model (b), Recent Migrants only (c), Model with College Indicator only (d), Model Excluding States with High Illegal Migration (e), and Model Excluding Bordering States (f).



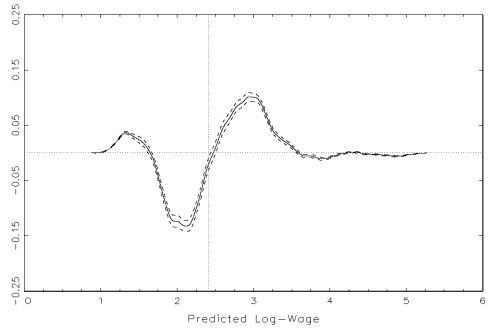
(a) Full Model (b) Parametric Model

(c) Recent Migrants (d) College indicator only

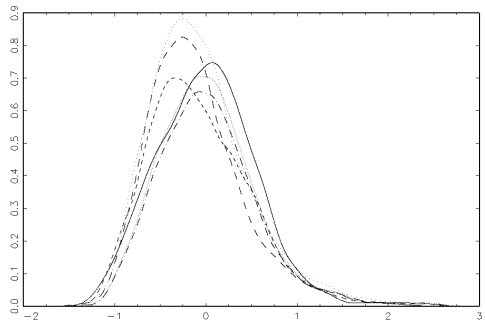
(e) Model excluding High Illegal Migration States (f) Model excluding Mexican-U.S. Bordering States

The Actual distribution represents the distribution of wages for Mexican-born workers currently residing in the U.S. The Counterfactual distribution represents the distribution of wages in the U.S. for Mexican-born workers if all migrants settled permanently, i.e., if no return migration occurred between 1995 and 2000. Section 4 in the paper explains how to recover these densities. Tables for the deciles of these distributions are available upon request. Standard errors have been bootstrapped (100 repetitions).
Model (a) estimates the distributions controlling also for length of stay in the U.S., industry and regional indicators. Column 2 of Table C.2 reports the regression results. Model (b) estimates the distribution using parametric techniques. Table C.3 reports the regression results of this estimation technique.
Model (e) reports the results excluding California, Florida, New York, and Texas.
Model (f) reports the results excluding the states on the Mexico-U.S. border.
KS test statistic for equality in the two distribution in model (a): 3.10, in model (b): 5.43, (c): 1.98, (d): 2.16, (e): 4.72, in model (f): 4.89. Critical value at 1%: 1.63.

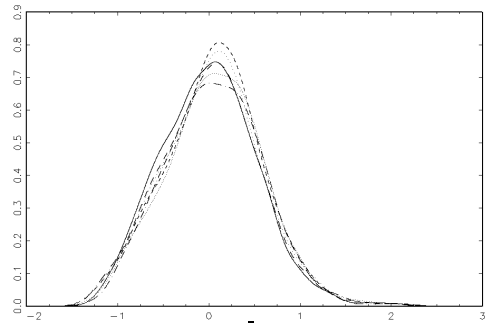
Figure 6: Counterfactual Distributions conditioning: Suggestive Evidence on Identifying Assumptions.



(a) Differences in Counterfactual and Actual Log-Wage Densities, without exclusion restriction



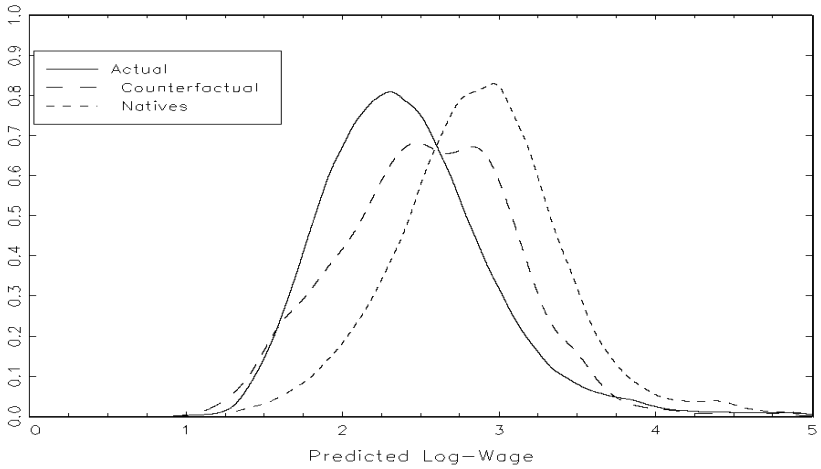
(b) Distribution of the residuals in the selected sample, conditional on different deciles of $Z_i'\hat{\alpha}$



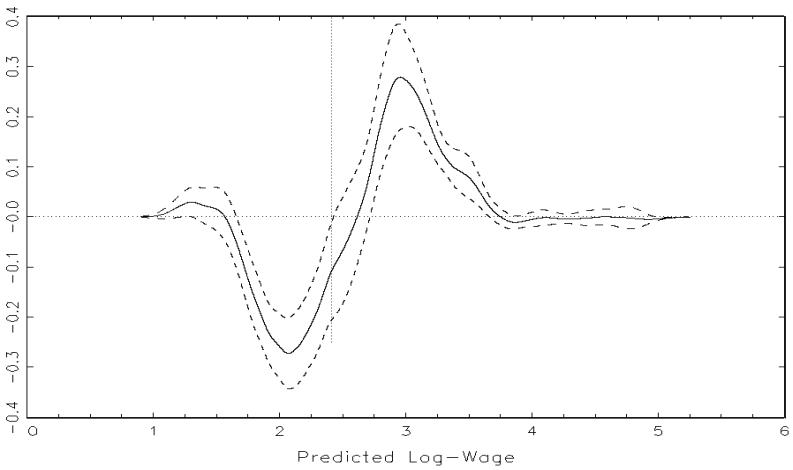
(c) Distribution of the residuals in the high probability set, conditional on different deciles of $Z_i'\hat{\alpha}$

Panel (a) reports the difference in the Counterfactual and Actual distributions in a parametric model without exclusion restrictions.
Panel (b) and (c) compare the distribution of u^* for individuals in various deciles $[q]$ of the $(Z_i'\alpha)$-index $(f(u^*|Z_i'\hat{\alpha}_{[q]}))$. If $Z_i'\hat{\alpha}$ was exogenous, these conditional distributions should be close.
Table 5 reports the Kolmogorov-Smirnov test for equality of the $f(u^*)$ and $f(u^*|Z_i'\hat{\alpha})$ distribution functions.

Figure 7: Actual and Counterfactual Log-Wage Distributions (a) and their Difference (b), conditioning on the 9th decile of the $(Z_i'\alpha)$-index



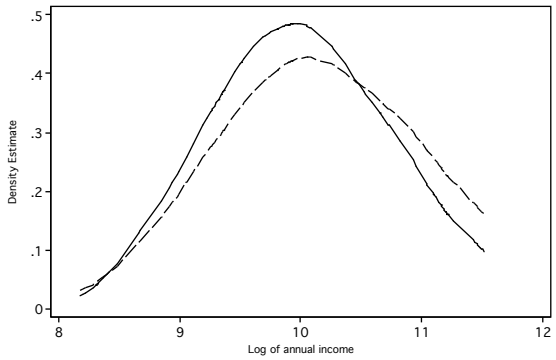
(a) Actual (Solid Line) and Counterfactual (Dashed Line) Wage Distribution



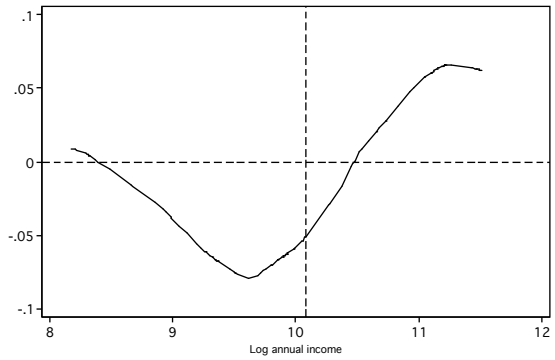
(b) Difference in Counterfactual and Actual Distributions

The Actual distribution represents the distribution of wages for Mexican-born workers currently residing in the U.S. The Counterfactual distribution represents the distribution of wages in the U.S. for Mexican-born workers, if all migrants settled permanent, i.e., if no return migration occurred between 1995 and 2000, conditioning on the 9th decile of the $(Z_i'\alpha)$-index. Standard errors have been bootstrapped (100 repetitions).
Kolmogorov-Smirnov test statistic for equality in the Actual and Counterfactual distributions: 3.80. Critical value at 1%: 1.63.

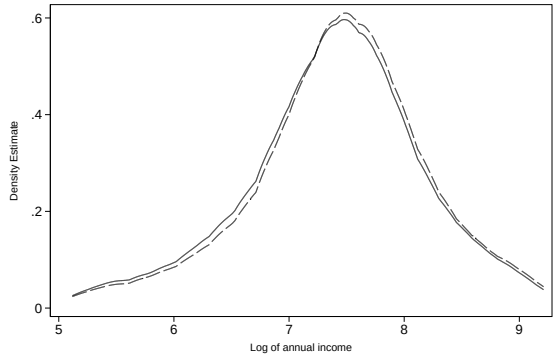
Figure 8: Actual Log-Income Distributions Mexican Family Life Survey and Migration Supplement, Mexican Census



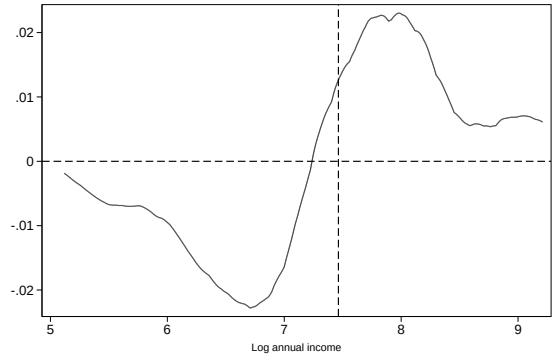
(a) Actual (Solid Line)Log- Income Distribution for Stayers ($S = 1$) and for all Migrants (Dashed Line), prior Migration to the US, MxFLS



(b) Difference in the two Distributions, MxFLS



(c) Income Distribution for Households with U.S. Stayers (Solid Line) and for all Households with Stayers and Returnees (Dashed Line), Mexican Census



(d) Difference in the two Distributions, Mexican Census

Source Panel (a) and (b): Mexican Family Life Survey. See text for explanation.
Source Panel (c) and (d): Mexican Census, Migration supplement. See text for explanation.
Notes: In panel (a), the Actual distribution represents the distribution of income before migration to the U.S. for Mexican-born workers currently residing in the U.S. The Counterfactual distribution represents the distribution of income prior migration of all Mexican-born workers who migrated to the U.S. - including the returnees. Panel (c) shows the log-income distribution of Mexican households where all their migrant members are still residing in the U.S. (solid line) and the log-income distribution of all Mexican households, with family members currently in the U.S. or having returned from the U.S. (counterfactual, dashed line).

References

Albrecht, J., A. Van Vuuren, and S. Vroman (2009). Counterfactual distributions with sample selection adjustments: econometric theory and an application to the Netherlands. *Labour Economics 16*(4), 383–396.

Ambrosini, J. W., K. Mayr, G. Peri, and D. Radu (2015). The selection of migrants and returnees in Romania: evidence and long-run implications. *Economics of Transition 23*(4), 753–793.

Ambrosini, J. W. and G. Peri (2012, February). The determinants and the selection of Mexico-US migrants. *The World Economy 35*(2), 111–151.

Andrews, D. W. K. and M. M. Schafgans (1998). Semiparametric estimation of the intercept of a sample selection model. *The Review of Economic Studies 65*(3), 497–517.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica 75*(2), 323–363.

Borjas, G. J. (1980). The relationship between wages and weekly hours of work: the role of division bias. *The Journal of Human Resources 15*(3), 409–423.

Borjas, G. J. (1985). Assimilation, changes in cohort quality and the earnings of immigrants. *Journal of Labor Economics 3*(4), 463–489.

Borjas, G. J. (1987). Self-selection and the earnings of immigrants. *American Economic Review 77*, 531–553.

Borjas, G. J. (1994, December). The economics of immigration. *Journal of Economic Literature 32*(4), 1667–1717.

Borjas, G. J. and B. Bratsberg (1996). Who leaves? The outmigration of the foreign-born. *The Review of Economics and Statistics 78*(1), 165–176.

Butcher, K. F. and J. DiNardo (2002). The immigrant and native-born wage distributions: evidence from United States censuses. *Industrial and Labor Relations Review 56*(1), 97–121.

Caponi, V. (2011). Intergenerational transmission of abilities and self-selection of Mexican immigrants. *International Economic Review 52*(2), 523–547.

Card, D. and E. G. Lewis (2007). The diffusion of Mexican immigrants during the 1990s: explanations and impacts. In *Mexican immigration to the United States*, pp. 193–228. University of Chicago Press.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics 32*(2), 189 – 218.

Chiquiar, D. and G. H. Hanson (2005). International migration, self-selection, and the distribution of wages: evidence from Mexico and the United States. *Journal of Political Economy 113*(2), 239–278.

Chiswick, B. R. (1978, October). The effect of Americanization on the earnings of foreign-born men. *The Journal of Political Economy 86*(5), 897–921.

Chzhen, Y. and K. Mumford (2011). Gender gaps across the earnings distribution for full-time employees in Britain: allowing for sample selection. *Labour Economics 18*(6), 837–844.

Delgado, M. A. and J. Mora (1995). Nonparametric and semiparametric estimation with discrete regressors. *Econometrica 63*(6), 1477–1484.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica 64*(5), 1001–1044.

Dos Santos, M. D. and F. Postel-Vinay (2003). Migration as a source of growth: the perspective of a developing country. *Journal of Population Economics 16*, 161–175.

Dustmann, C. (2003). Return migration, wage differentials and the optimal migration duration. *Europan Economic Review 47*, 353 – 369.

Dustmann, C., I. Fadlon, and Y. Weiss (2011). Return migration, human capital accumulation and the brain drain. *Journal of Development Economics 25*(1), 58–67.

Dustmann, C. and Y. Weiss (2007). Return migration: theory and empirical evidence from the UK. *British Journal of Industrial Relations 45*(2), 236–256.

Fernandez-Huertas Moraga, J. (2011). New evidence on emigrant selection. *The Review of Economics and Statistics 93*(1), 72–96.

Hanson, G. H. (2006, December). Illegal migration from Mexico to the United States. *Journal of Economic Literature 44*, 869–924.

Heckman, J. J. (1990). Varieties of selection bias. *The American Economic Review 80*(2), 313–138, Papers and Proceedings.

Hu, W.-Y. (2000). Immigrant earnings assimilation: estimates from longitudinal bata. *The American Economic Review 90*(2), 368–372.

Huber, M. and G. Mellace (2014). Testing exclusion restrictions and additive separability in sample selection models. *Empirical Economics 47*(1), 75–92.

Ibarraran, P. and D. Lubotsky (2007). Mexican immigration and self-selection: new evidence from the 2000 Mexican census. In *Mexican Immigration to the United States*, NBER Chapters, pp. 159–192. National Bureau of Economic Research, Inc.

Kaestner, R. and O. Malamud (2014). Self-selection and international migration: new evidence from Mexico. *Review of Economics and Statistics 96*(1), 78–91.

Kitagawa, T. (2010). Testing for instrument independence in the selection model. mimeo.

Klein, R. and C. Shen (2010). Bias corrections in testing and estimating semiparametric, single index models. *Econometric Theory 26*(06), 1683–1718.

Klein, R., C. Shen, and F. Vella (2015). Estimation of marginal effects in semiparametric selection models with binary outcomes. *Journal of Econometrics 185*(1), 82–94.

Klein, R. W. and R. H. Spady (1993, March). An efficient semiparametric estimator for binary response models. *Econometrica 61*(2), 387–421.

Lacuesta, A. (2010). A revision of the self-selection of migrants using returning migrants earnings. *Annals of Economics and Statistics 97/98*, 235–259.

LaLonde, R. J. and R. H. Topel (1992). The assimilation of immigrants in the us labor market. In *Immigration and the workforce: economic consequences for the United States and source areas*, pp. 67–92. University of Chicago Press.

Lanot, G. and I. Walker (1998). The union/non-union wage differential: an application of semi-parametric methods. *Journal of Econometrics 84*(2), 327–349.

Lindstrom, D. P. and D. S. Massey (1994). Selective emigration, cohort quality, and models of immigrant assimilation. *Social Science Research 23*(4), 315 – 349.

Liu, E., C. Hsiao, T. Matsumoto, and S. Chou (2009). Maternal full-time employment and overweight children: parametric, semi-parametric, and non-parametric assessment. *Journal of Econometrics 152*(1), 61–69.

Lubotsky, D. (2007). Chutes or ladders? A longitudinal analysis of immigrant earnings. *Journal of Political Economy 115*(5), 820–867.

Martins, M. F. O. (2001). Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal. *Journal of Applied Econometrics 16*(1), 23–39.

Massey, D. S. and K. E. Espinosa (1997). What's driving Mexico-U.S. migration? A theoretical, empirical, and policy analysis. *The American Journal of Sociology 102*(4), 939–999.

McKenzie, D. and H. Rapoport (2010). Self-selection patterns in Mexico-U.S. migration: the role of migration networks. *The Review of Economics and Statistics 92*(4), 811–821.

Mishra, P. (2007). Emigration and wages in source countries: evidence from Mexico. *Journal of Development Economics 82*(1), 180–199.

Mulligan, C. and Y. Rubinstein (2008). Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics 123*(3), 1061–1110.

Passel, J. S. (2006). The size and characteristics of the unauthorized migrant population in the U.S. Research report, Pew Hispanic Center.

Reinhold, S. and K. Thom (2013). Temporary migration, skill upgrading, and legal status: evidence from Mexican migrants. *Journal of Human Resources 48*(3), 768–820.

Robinson, P. (1988, July). Root-N bonsistent semiparametric regression. *Econometrica 56*(4), 931–954.

Schafgans, M. M. (2000). Gender wage differences in Malaysia: parametric and semiparametric estimation. *Journal of Development Economics 63*(2), 351–378.

Schafgans, M. M. A. (1998). Ethnic wage differences in Malaysia: parametric and semiparametric estimation of the chinese-Malay wage gap. *Journal of Applied Econometrics 13*(5), 481–504.

Schafgans, M. M. A. and V. Zinde-Walsh (2002, February). On the intercept estimation in the sample selection model. *Econometric Theory 18*(1), 40–50.

Shen, C. (2013). Determinants of health care decisions: insurance, utilization, and expenditures. *Review of Economics and Statistics 95*(1), 142–153.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Yang, D. (2006, November). Why do migrants return to poor countries? Evidence from Philippine migrants' responses to exchange rate shocks. *Review of Economic and Statistics 88*(4), 715–735.

A Monte Carlo

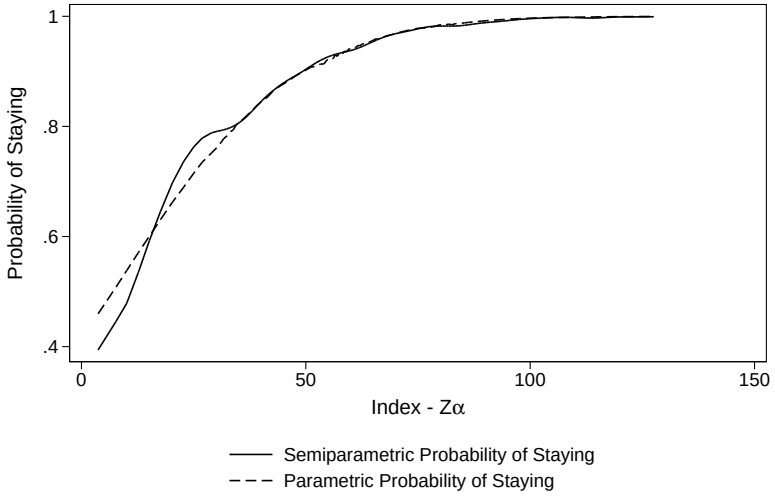
To get some sense of how well the presented method works, I conducted a small Monte Carlo experiment. The data generating process is the following:

$$S_i = \begin{cases} 1 & c + Age_i + Primary_i + Secondary_i + College_i - Married_i + USbornSpouse_i + Child_i - USbornChild_i > \epsilon_i \\ 0 & c + Age_i + Primary_i + Secondary_i + College_i - Married_i + USbornSpouse_i + Child_i - USbornChild_i \leq \epsilon_i \end{cases}$$

$$Y_i = 1 + Age_i + Primary_i + Secondary_i + College_i + Married_i + USbornSpouse_i + Child_i + u_i \qquad \text{if} \quad S_i = 1$$

I let the error term being distributed as a standard normal. This is a simple assumption not needed in the Klein and Spady estimator, which is free of distributional assumptions and can also handle heteroskedasticity of a general but known form, or of an unknown form with an index structure. As the marginal effects from the baseline model (Table 1) are close to the marginal effects of a probit model (Table C.3) and similarly results from the full semiparametric procedure (Figure 3) are similar to results from a parametric estimation (Figure 5(b)), normality seems a good approximation of the data generating process. This can be seen also in Figure 9 below, which plots the semiparametric and parametric responses, as a function of the selection index. The two response functions are very similar.

Figure 9: Semiparametric estimation of the Probability of Staying in the U.S., Men, 25-55 Years Old.



The figure plots the semiparametric and parametric response functions, as a function of the selection index.

In the data generating process, all parameters are equal to 1 in absolute value, while their signs are set to reflect the results of the empirical specification (and common sense). The constant c has been adjusted to match the share of stayers in the generated data with that of the sample. All variables are discrete, with the exception of Age. Age takes 31 distinct values as in the empirical application. Table A.1 shows the means of all the generated variables: the first moments of the generated sample closely resemble the first moments of the variables in the data.

Table A.1: Comparison of Empirical and Generated Means

	Data	Monte Carlo
Stayers	0.90	0.90
Age	35.98	36.07
Primary	0.47	0.47
Secondary	0.32	0.32
College	0.04	0.04
Married	0.80	0.80
US born Spouse	0.10	0.11
Child	0.63	0.64
US born Child	0.57	0.57

Means of all generated variable across all replications (R=1,000). Sample size set to 5,000 observations.

The Monte Carlo experiment is run on the estimator proposed in section 4.1.[29] For each iteration in the Monte Carlo experiment, I calculate the deciles of the distribution of u^*, estimated as explained in the paper, and the deciles of the distribution of u^* for those observations for which $S_i = 1$, i.e. for the stayers, and for the observations in the high probability set. These represent the deciles of the two distributions of interest: the 'actual' distribution, $\hat{f}(u^*|S_i = 1)$, and the counterfactual distribution, $\hat{f}(u^*)$. Due to sample selection, the deciles of the actual distribution should be far from the deciles of the normally distributed random variable u^*, while, if the estimator proposed in equation (4) works, the deciles of the distribution in the high probability set should be close to the deciles of a normal distribution. I run this experiment for $N = 5,000$, $N = 10,000$ and $N = 60,000$ with 1,000 replications each. Table A.2 reports the bias between each decile of $\widehat{f}(u^*|S_i = 1)$ or $\widehat{f}(u^*)$ and a normally distributed random variable. The first, third and fifth columns of the table shows how using the distribution of the error term in the selected sample does not recover the true distribution in the population: in fact, the estimation of each decile of the distribution is consistently biased. On the contrary, column two, four and six reports the deciles of the distribution estimated using (4). Across all sample sizes, the estimator performs very well and the bias is negligible. This suggests that the estimator in equation (4) is able to recover the true distribution in the presence of self-selection.

[29] The small and large sample properties of the Klein and Spady's estimator as well as the Robinson's estimator have been shown elsewhere. In this section, instead, I assume that good estimates of the parameters are available to the researcher and run a Monte Carlo that instead focuses on the estimation of the counterfactual distribution in the high probability set - the key advance of the paper. I have run a complementary small Monte Carlo experiment where I put on scrutiny the performance of the full estimation technique. Results are available upon request.

Table A.2: Comparison of the Deciles of $\hat{f}(u_i^*)$ and $\hat{f}(u_i^*|S_i = 1)$ with the Deciles of a Normal Random Variable.

	N = 5,000		N = 10,000		N = 60,000	
Decile	$f(u^*\|S=1)$	$f(u^*)$	$f(u^*\|S=1)$	$f(u^*)$	$f(u^*\|S=1)$	$f(u^*)$
1.0	-0.469	-0.018	-0.469	-0.012	-0.467	-0.009
2.0	-0.494	-0.017	-0.494	-0.014	-0.493	-0.011
3.0	-0.514	-0.015	-0.513	-0.012	-0.513	-0.012
4.0	-0.530	-0.019	-0.530	-0.017	-0.529	-0.014
5.0	-0.545	-0.025	-0.545	-0.020	-0.545	-0.017
6.0	-0.562	-0.028	-0.562	-0.026	-0.561	-0.022
7.0	-0.579	-0.036	-0.579	-0.033	-0.579	-0.029
8.0	-0.601	-0.049	-0.601	-0.044	-0.600	-0.040
9.0	-0.631	-0.080	-0.632	-0.071	-0.629	-0.069

B Accounting for Heteroskedasticity

Suppose that the model is:

$$Y_i^* = X_i'\beta_0 + c_0 + e_i^*,$$

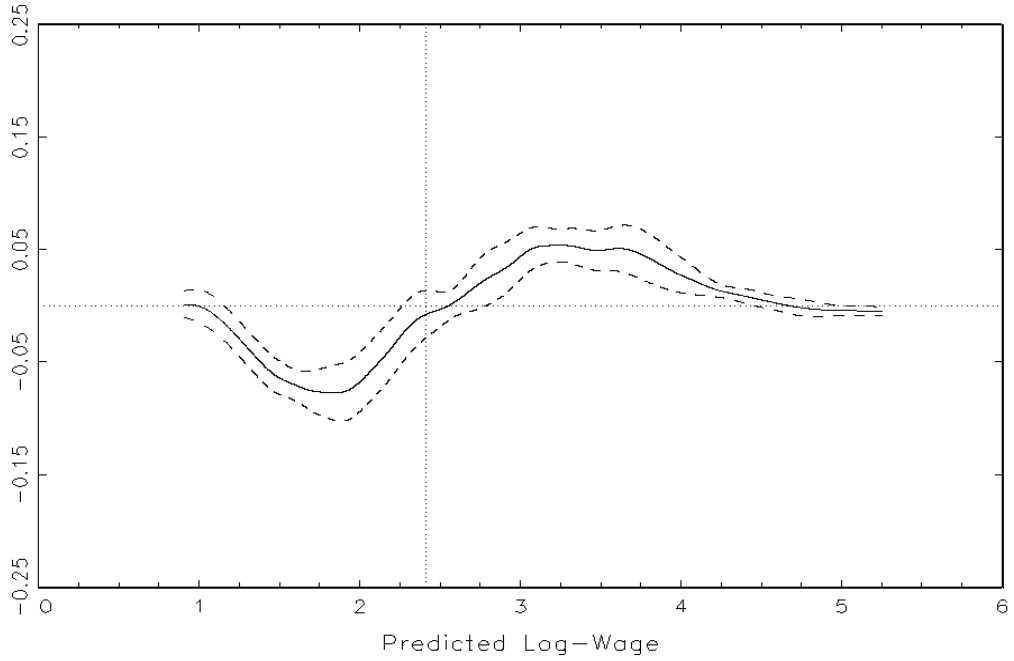
where there is heteroskedasticity in e^* of unknown form, i.e. $e_i^* = u^* k(X\delta_0)$. The observed model could be written as:

$$Y_i = X_i'\beta_0 + c_0 + G(Z_i'\alpha_0) + u^* k(X\delta_0),$$

where $G(\cdot)$ is the piece due to selection and $k(X\delta_0)$ is the piece due to heteroskedasticity.

To allow for heteroskedasticity, the following estimation strategy was introduced. As in the high probability set $G(Z_i'\alpha_0)$ tends to zero, it is possible in this set to estimate semiparametrically $\hat{k}(\cdot)$ simply by estimating the conditional variance of the model. A simple GLS estimator recovers then the parameters δ. I report below then the difference in the standardized distribution of u^* in the selected sample of stayers and in that of returnees. As it can be seen the main conclusions of the paper are still obtained.

Figure 10: Difference in Counterfactual and Actual Distributions accounting for Heteroskedasticity



Standard errors have been bootstrapped (100 replications).

C Additional Tables

Table C.3: Demographic and socio-economic characteristics, Native Born and Foreign Born Men, 25-55 Years Old

Variable	**Natives**	**All Mexican Born**	**Stayers**	**Returnees**
Age	39.697	35.977	36.168	34.237***
	(8.483)	(7.915)	(7.935)	(7.508)
Less than Primary School	0.003	0.165	0.155	0.258***
	(0.056)	(0.372)	(0.362)	(0.437)
Primary Education	0.068	0.468	0.455	0.589***
	(0.251)	(0.499)	(0.498)	(0.492)
Secondary Education	0.647	0.324	0.347	0.112***
	(0.478)	(0.468)	(0.476)	(0.315)
College Education	0.282	0.043	0.043	0.042
	(0.450)	(0.203)	(0.204)	(0.201)
Married	0.806	0.796	0.796	0.795
	(0.395)	(0.403)	(0.403)	(0.404)
US born spouse	0.677	0.105	0.115	0.011***
	(0.468)	(0.306)	(0.319)	(0.104)
Child	0.528	0.629	0.630	0.623
	(0.499)	(0.483)	(0.483)	(0.485)
US born child	0.534	0.568	0.611	0.179***
	(0.499)	(0.495)	(0.488)	(0.384)
0-5 Years in U.S.	-	-	0.161	-
	-	-	(0.367)	-
5-10 Years in U.S.	-	-	0.167	-
	-	-	(0.373)	-
10-20 Years in U.S.	-	-	0.387	-
	-	-	(0.487)	-
20-30 Years in U.S.	-	-	0.226	-
	-	-	(0.418)	-
30-40 Years in U.S.	-	-	0.049	-
	-	-	(0.215)	-
>40 Years in U.S.	-	-	0.011	-
	-	-	(0.103)	-

Continue to next page

Continued from previous page

Variable	Natives	All Mexican Born	Stayers	Returnees
Northeast Region	0.188	-	0.028	-
	(0.391)	-	(0.165)	-
Midwest	0.255	-	0.103	-
	(0.436)	-	(0.304)	-
South	0.359	-	0.281	-
	(0.480)	-	(0.450)	-
West	0.198	-	0.588	-
	(0.398)	-	(0.492)	-
Agriculture, fishing, and forestry	0.023	0.147	0.136	0.245***
	(0.149)	(0.354)	(0.343)	(0.430)
Mining	0.010	0.006	0.006	0.006
	(0.099)	(0.075)	(0.074)	(0.080)
Manufacturing	0.220	0.219	0.231	0.102***
	(0.414)	(0.413)	(0.422)	(0.302)
Electricity, gas and water	0.019	0.004	0.004	0.002***
	(0.135)	(0.064)	(0.066)	(0.040)
Construction	0.114	0.195	0.204	0.115***
	(0.318)	(0.397)	(0.403)	(0.319)
Wholesale and retail trade	0.168	0.191	0.202	0.086***
	(0.374)	(0.393)	(0.402)	(0.280)
Hotels and restaurants	0.006	0.017	0.015	0.032***
	(0.080)	(0.129)	(0.123)	(0.176)
Transportation and Communications	0.073	0.039	0.038	0.047***
	(0.260)	(0.194)	(0.192)	(0.212)

Continue to next page

Continued from previous page

Variable	Natives	All Mexican Born	Stayers	Returnees
Financial services	0.036	0.005	0.006	0.003***
	(0.187)	(0.074)	(0.075)	(0.054)
Public administration and defense	0.075	0.011	0.011	0.016***
	(0.264)	(0.106)	(0.104)	(0.126)
Real estate and business services	0.080	0.062	0.067	0.017***
	(0.272)	(0.241)	(0.250)	(0.131)
Education	0.054	0.015	0.015	0.010***
	(0.226)	(0.121)	(0.123)	(0.101)
Heath and social work	0.045	0.012	0.013	0.004***
	(0.208)	(0.108)	(0.112)	(0.061)
Other services	0.077	0.050	0.050	0.046**
	(0.266)	(0.217)	(0.218)	(0.209)
Private household services	0.000	0.002	0.001	0.005***
	(0.019)	(0.039)	(0.035)	(0.067)
Wage	21.555	13.432	13.432	-
	(17.200)	(11.585)	(11.585)	-
Observations	103,994	133,389	120,205	13,184

Standard deviations in parentheses
Significance levels: *: 10%, **: 5%, ***: 1% for a t-test for differences in means between Returnees and U.S. Stayers.

Table C.2: Wage Equation Estimates, Mexican-Born Men Working for Wages, 25-55 Years Old.

	(1)	(2)
Constant	1.658***	1.602***
	(0.279)	(0.259)
Age	0.026***	0.021***
	(0.003)	(0.003)
Age Sq	-2.532E-04***	-2.503E-04***
	(3.390E-05)	(3.350E-05)
Primary Education	0.038***	0.022***
	(0.005)	(0.005)
Secondary Education	0.156***	0.118***
	(0.006)	(0.005)
College Education	0.461***	0.403***
	(0.008)	(0.008)
Married	0.039***	0.045***
	(0.004)	(0.004)
US born spouse	0.063***	0.041***
	(0.007)	(0.006)
Child	0.118***	0.095***
	(0.004)	(0.004)
5-10 Years in U.S.	-	0.043***
	-	(0.005)
10-20 Years in U.S.	-	0.117***
	-	(0.005)
20-30 Years in U.S.	-	0.210***
	-	(0.006)
30-40 Years in U.S.	-	0.299***
	-	(0.009)
>40 Years in U.S.	-	0.393***
	-	(0.017)
Industry indicators	No	Yes
Regional indicators	No	Yes
R^2	0.069	0.123
R^2-adjusted	0.069	0.123
N	120,205	120,205

Standard errors in parentheses
Significance levels: *: 10%, **: 5%, ***: 1%.
The industry and regional indicators used in column (3) and (4) are the variables presented in the descriptive statistics.

Table C.3: Probit and Wage Equation Estimates, Parametric Model, Men working for wages, 35-55 Years old.

	Probit Marginal Effects, $S = 1$	Wage Equation
Baseline	0.948	
Constant	-	1.558***
	-	(0.056)
Age	0.003***	0.022***
	(0.001)	(0.002)
Age Sq	9.340E-06	-2.6E-04***
	(1.000E-05)	(2.4E-05)
Primary Education	0.020***	0.024***
	(0.001)	(0.004)
Secondary Education	0.076***	0.129***
	(0.001)	(0.005)
College Education	0.026***	0.404***
	(0.002)	(0.008)
Married	-0.006***	0.043***
	(0.002)	(0.004)
US born spouse	0.051***	0.057***
	(0.001)	(0.005)
Child	-0.073***	0.100***
	(0.001)	(0.004)
US born child	0.186***	-
	(0.002)	-

Continue to next page

Continued from previous page

	Probit Marginal Effects, $S = 1$	Wage Equation
5-10 Years in U.S.		0.044***
		(0.005)
10-20 Years in U.S.		0.120***
		(0.005)
20-30 Years in U.S.		0.214***
		(0.005)
30-40 Years in U.S.		0.305***
		(0.008)
>40 Years in U.S.		0.390***
		(0.016)
Lambda		-0.061***
		(0.007)
Industry indicators	No	Yes
Regional indicators	No	Yes
N	133,389	120,205

Standard errors in parentheses
Significance levels: *: 10%, **: 5%, ***: 1%.
The industry and regional indicators used in column (3) and (4) are the variables presented in the descriptive statistics.