DISCUSSION PAPER SERIES

IZA

IZA DP No. 6795

Recovering the Counterfactual Wage Distribution with Selective Return Migration

Costanza Biavaschi

August 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Recovering the Counterfactual Wage Distribution with Selective Return Migration

Costanza Biavaschi
IZA

Discussion Paper No. 6795
August 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 6795
August 2012

ABSTRACT

Recovering the Counterfactual Wage Distribution with Selective Return Migration

This paper explores the distribution of immigrant wages in the absence of return migration from the host country. In particular, it recovers the counterfactual wage distribution if all Mexican immigrants were to settle in the United States and no out-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 accounts also for selection on unobservables. I adopt a semiparametric procedure that recovers this counterfactual distribution and find that Mexican returnees are middle- to high-wage earners at all levels of educational attainment. The presented results contrast with the general perception that those migrants who return home have failed in the host country.

JEL Classification: J61, F22

Keywords: return migration, self-selection, assimilation, U.S.-Mexico migration

Corresponding author:

Costanza Biavaschi
IZA
P.O. Box 7240
53072 Bonn
Germany
E-mail: biavaschi@iza.org

1 Introduction

Considering that the United States continues to be the primary destination for international migrants (Özden et al., 2011), it is not surprising that the immigration debate has often been pivotal in its political and academic arena. Indeed, a vast literature has already investigated the key questions about cross-border migration, such as why people migrate, how migrants fare in the host country, and the impact of immigration on the native-born population of the host country. However, previous research has tended to overlook that a proportion of immigrants return to their home countries. Taking into account such return migration urges scholars to reconsider how they model migration decisions as well as how they measure the effects of migration on both immigrants and natives and on both the sending and the receiving regions.

In light of the foregoing, this paper combines data derived from the U.S. and Mexican censuses of 2000 to estimate the wage distributions of Mexican immigrants in the U.S. under two conditions – with and without return migration. This approach enables us to answer two main questions. First, how do returnees compare with stayers and where does return migration have its largest impact on the wage distribution? Second, how would the immigrant wage distribution change in the host economy if incentives to return were altered based on exogenous variations in economic opportunities in the source or in the host countries? For example, there is evidence that during periods of financial turmoil some migrants respond to exchange rate shocks by returning home, whereas others reduce their return intentions (Yang, 2006). Moreover, inflows and outflows are also typically influenced by migration policies, such as enforcement by border patrols (Thom, 2010; Angelucci, 2012).[1] This paper thus highlights the consequences for the U.S. if no return migration of Mexican-born workers had occurred between 1995 and 2000.

Several important contributions to the body of knowledge on this topic have examined how return migration influences the average wage levels of immigrants (Hu, 2000; Lubotsky, 2007), especially Mexican immigrants (Lindstrom and Massey, 1994; Reinhold and Thom, 2009; Lacuesta, 2010). The major challenges faced by of all these studies, however, are the lack of administrative data collected by immigration authorities and the fact that individuals self-select into the decision to return. Therefore, they all have certain limitations. For example, some have no actual information on return migrants but rather infer return migration from survey samples (Hu, 2000; Lubotsky, 2007), some fail to use representative samples (Lindstrom and Massey, 1994; Reinhold and Thom, 2009), while others argue against the importance of selectivity and compare Mexicans in Mexico with and without experience in the U.S. (Lacuesta, 2010).

By contrast, the present paper not only uses representative data but also accounts for selection based on both the observable and the unobservable attributes of migrants, focusing on the whole wage distribution. Because migrants' return is an individual decision, the overarching problem is to recover the counterfactual wage density in the presence of selective return migration. This paper thus adopts a semiparametric procedure in the spirit of Heckman (1990) that complements the estimator used by Chiquiar and Hanson (2005), which accounted for selection based on observable traits. Such an estimation strategy provides an alternative to using pre-migration earnings to measure selectivity (see Kaestner and Malamud (2010); Fernandez-Huertas Moraga (2011); Ambrosini and Peri (2012); McKenzie and Rapoport (2010)), as such

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

information is often either unavailable to researchers or the return flow in these surveys is too small to allow suitable analysis (Ambrosini and Peri, 2012).

I find that, conditioning on observable characteristics, Mexican returnees are middle to high wage earners, consistent with models in which the decision to return hinges on reaching desired goals in the host country. Overall, the return flow has a small effect on immigrant wage inequality: the outflow of immigrants decreases dispersion in the lower part of the distribution and it increases it in the upper part. 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 in designing optimal migration policies policy makers 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 slightly close 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 Section 7 shows the robustness of the analysis. Conclusions are drawn in Section 8.

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

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

First, the results of Chiquiar and Hanson's (2005) study of the selection of immigrants contradict the theoretical predictions put forward in Borjas (1987), as they showed intermediate to positive selection based on the observable characteristics of Mexican immigrants to the U.S. compared with Mexican stayers in Mexico. By contrast, using only the information provided in the Mexican census about experience in the U.S., Ibarraran and Lubotsky (2007) find negative selection on observable skills. They also identify various possible reasons for these inconsistent results, including the mistranslation or misunderstanding of the grade and degree choices in the U.S. census (which may, in turn, cause the misreporting of the education variable-the key factor in studying selection) and the undercount of young and largely illegal Mexican immigrants in the U.S. census.

Moreover, Fernandez-Huertas Moraga (2011), using the *Encuesta Nacional de Empleo Trimestral*, quantifies the source of the discrepancies between two previous studies, showing that the results presented by Chiquiar and Hanson (2005) are primarily driven by the undercount of unskilled immigrants in the U.S. as well as by the omission of unobservables in the estimation procedure. In addition, McKenzie and Rapoport (2010) use the *Encuesta Nacional de la Dinámica Demográfica* in order to reconcile Borjas's (1987) theoretical framework with Chiquiar and Hanson's (2005)) empirical findings, suggesting that the results are driven by the differential impact of education in communities that have large and small networks. Based on information derived from the Mexican Family Life Survey, Kaestner and Malamud (2010) suggest that migrants are selected from the middle of the education and wage distributions, although after controlling for network effects these results are not as strong. Interestingly, the authors also find little evidence of selection on unobservables. Using the same data source, Ambrosini and Peri (2012) confirm the findings of Borjas (1987) and Fernandez-Huertas Moraga (2011) on the negative selection of Mexican immigrants to the U.S., a result primarily driven by differences in the unobservable skills of migrants and non-migrants.

In summary, the current debate has 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 (Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012; McKenzie and Rapoport, 2010). Second, researchers must aim to control for the unobservable differences between migrants and non-migrants (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012).

Turning now 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 by 10% to 15% the wage progress of stayers. Likewise, Hu (2000) shows a decline in immigrant wage growth once return migration has been taken into account. Such results are, however, 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 in their longitudinal datasets returnees are not directly identified and return migration cannot be separated from other sources of panel attrition.[2]

Lacuesta (2010) and Reinhold and Thom (2009) both recently 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 (2009), 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.

The foregoing confirms that self-selection and data availability have impeded a full understanding of return migration and its consequences. 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 the U.S. and Mexican censuses. This approach allows researchers to distinguish return migration from panel attrition and treat all those forms of sample selection and heterogeneity that are not eliminated by fixed effects estimators in panel data analyses. Further, it provides a full picture of what the U.S. could expect if return migration were zero, because of changes in either migration policies or migration incentives.

On the methodological side, the present paper also introduces an estimator for a counterfactual dis-

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

tribution that accounts for sample selection. This technique complements the analysis based on selection on observables (Chiquiar and Hanson (2005)[3], Ibarraran and Lubotsky (2007)) 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 presented analysis uses the Integrated Public Use Microdata Series: International of the U.S. and Mexican censuses from 2000. Mexican-born immigrants are defined as individuals born in Mexico that appear in the U.S. census, while Mexican-born return migrants are defined as temporary migrants in the U.S. that appear in the Mexican census, with returnees identified as those who report having resided in the U.S. in the five years preceding the Mexican census. For comparison purposes, this study also uses data on a random sample of U.S. native-born workers (n = 103,994).

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 are changes in education once in the U.S., the misreporting of education in the U.S. census, and illegal immigration. Since this study focuses on return migration, the possibility that Mexican immigrants have obtained additional schooling after arriving in the U.S. should not be as invalidating, since returnees could have made the same choice. Nonetheless, there remains a concern that 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 in part be due to misreporting in the U.S. census rather than the selection of returnees. However, 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), and this could reduce such concerns. The undercount of illegal immigrants in the U.S. census might indeed constitute a problem (see Section 7). Lastly, there is a final worry specific to this study: the universe of returnees is much broader than that captured by the Mexican census. Since no further information is available on having been abroad, looking at place of residence in 1995 is the best proxy for return status. 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. Nevertheless, throughout the analysis this is assumed not to be the case; in other words, the sample is considered to be representative of the full population of returnees.[4]

The sample is restricted to men aged between 25 and 55 years, born in Mexico, and in earning

[3] This estimation is in turn based on DiNardo et al. (1996)

[4] 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.

employment. The total sample size is 133,389. Of these, 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 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*).[5] Throughout this analysis, the decision to stay in the U.S. is thus modeled as a function of these educational and socioeconomic variables.

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 (Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2012; McKenzie and Rapoport, 2010; Lacuesta, 2010). Moreover, the indicator of having a U.S.-born spouse is included in order to capture the constructs of "attachment"and "networks,"which have both been shown to be relevant in this type of analysis (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, but it 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 if they 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.: returnees thus seem to be negatively selected in terms of observable characteristics. Consequently, the figure suggests that, based on observable traits, returnees are drawn disproportionately from the bottom of the wage distribution. Section 4 explains how a counterfactual distribution if all returnees had stayed can be estimated by accounting both for observable and for unobservable traits. The remainder of the paper then compares these results with the descriptive analysis presented in this section.

4 The Model and the Estimation Strategy

The research question answered in this study requires the recovery of the wage distribution for all Mexican-born men who have been in the U.S., even though wages are observed only for Mexican-born immigrants

[5] All other average characteristics for the sample are reported in Table C.2 in the Appendix. 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.

who are currently residing in the country. Let S_i be an indicator of whether or not 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:

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

Let the true wage determination process for a randomly selected Mexican immigrant present in the U.S. be:

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

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 the 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 reminder of the paper focuses on estimation techniques that are free of 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. Section 7 discusses these assumptions. Subsection 4.1 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, $(X_i'\beta_0 + c_0)$. Most of the following discussion will therefore focus 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.:

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

Under a stronger assumption of independence, $f(u_i^*) = f(u_i^*|Z_i'\alpha_0)$. The analysis is carried under this assumption for computational speed and expositional purposes. Independence is, however, not necessary.

Results are similar when conditioning on particular quantiles of the selection index, as further discussed in Section 7.

This density cannot be directly estimated using the sample wage distribution, as the latter is observed only 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 $f(u_i^*|S_i = 0, Z_i'\alpha_0)$. However, note that whenever $\Pr(S_i = 1|Z_i'\alpha_0)$ is close to 1, $f(u_i^*) \approx f(u_i^*|S_i = 1, Z_i'\alpha_0)$. Intuitively, selection disappears in the limit for individuals for which $\Pr(S = 1|Z_i'\alpha_0)$ is close to one, namely for those individuals in a *high probability* set. This intuition is known as *identification at infinity* (Chamberlain, 1986), as advocated by Heckman (1990).[6] To my knowledge, such an identification strategy has not been applied to recover a counterfactual distribution, as 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. Throughout the paper, a Gaussian kernel with optimal bandwidth is chosen. 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 semiparametric dichotomous choice model,[7] by applying the estimation method developed by Klein and Spady (1993).[8]

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 $\Pr(\widehat{S_i = 1}|Z_i'\hat{\alpha})$.[9]

---

[6] 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. (2011) extend these results to allow for a definition of an high probability set that is data-dependent.

[7] 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.

[8] Effectively, in the estimation of the selection index, the only identified parameters in terms of the original model is the coefficient ratio, i.e., α_j/α_1, with $j = 1 \dots k$ and where α_1 is the coefficient of the continuous variable, which is normalized to 1. In order to reduce the notational burden, I disregard this technicality in the reminder of the discussion.

[9] 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.

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, there is 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.[10] 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.[11]

4.3 Potential and Limitations of the Estimation Strategy

We assume that $(X_i, Z_i, u_i^*, \epsilon_i)$ are i.i.d and that the regressors are exogenous. This has common aspects with the standard 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 (Kaestner and Malamud, 2010; Lacuesta, 2010; Reinhold and Thom, 2009), 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 assumed to be exogenous.[12] These models further conjecture the absence of an 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

[10] Additional regressions also controlled for the language spoken at home. Having a U.S. child might be related with 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.

[11] The research of valid exclusion restrictions poses many challenges to the migration area of study. I propose three methods to check the sensitivity of the results to the choice of exclusion restriction. First, I adopt a parametric estimation of these counterfactual densities, which is shown to yield similar results to the semiparametric procedure. Under the normality assumption, the parametric procedure has the advantage that identification can be reached through the non-linearities in the functional form of the selection term. Even when no variable is excluded from the model, the results presented in the paper are still recovered. Second, technically, identification at infinity does not require the use of an exclusion restriction (Chamberlain, 1986), but identification is reached through regressors having a larger probability mass at the tails compared with the error term (Andrews and Schafgans, 1998; Klein et al., 2011). Complementary models that do not use exclusion restrictions were run, and these provided similar results to those presented here. Third, 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 change at different quantiles of the estimated index $Z_i\hat{\alpha}$. In fact, an invalid exclusion restriction 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. As discussed in Section 7, Figure 5 shows the distribution in the high probability set conditional on the quantiles of $Z_i\hat{\alpha}$ and Table 4 reports the results of a Kolmogorov-Smirnov (KS) test of the differences in the estimated distribution of u^* and the conditional distribution $u^*|Z_i\hat{\alpha}$. The small variation in the two distributions is indicative of evidence that the distribution of the error term is not sensitive to changes in the index quantiles.

[12] 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.

nevertheless 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. These scenarios include whenever the data provide insufficient information on returnees' wages, such as in Lacuesta (2010) and Ambrosini et al. (2011), when the sample size is too small to guarantee enough 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 Kaestner and Malamud (2010), Fernandez-Huertas Moraga (2011), Ambrosini and Peri (2012), and McKenzie and Rapoport (2010).

5 Results

In addition to the interest in the counterfactual estimation, the data allow us to study 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 that determine 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 that have 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 that indicate social attachment to the host country are strongly significant, while-based on observable characteristics-stayers are more likely to have better educational outcomes.

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 compared with the various 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.

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; and (iii) how does this distribution change, conditional on educational characteristics?

How different is the immigrant population to the population of stayers in the U.S., in terms of observable and unobservable traits? Table 2 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. They are 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. The deciles of the predicted wage are reported as a summary measure.

[TABLE 2 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 quantiles 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 0.9%) 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 2. 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 presented evidence suggests that immigrant stayers and the full population (i.e., stayers and returnees) are somehow close in terms of observable traits, whereas differences arise in terms of unobservable traits. In particular, although returnees are a disadvantaged group in the labor market in terms of observable traits, their unobservable abilities seem to compensate for this lack of skills. Further, it seems as though 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.[13]

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 2. This panel 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 observable (panel one) and unobservable components (panel two) for the immigrant population at each decile. At almost all deciles, the implied counterfactual distribution 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 2(a) presents the actual and counterfactual distributions just 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 2(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 then positive. Therefore, the disadvantage in terms of lost human capital skills that returnees face is balanced by the higher unobserved motivation and productivity displayed by this group. This balance overall 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 2 HERE]

This finding is not the only insight from the analysis, however. The last panel in Table 2 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 8% 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, in the absence of return migration inequality within the Mexican population would increase slightly. 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. Table 3 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, the differences in observables are negligible across all educational groups, while unobservables are shown to drive the dissimilarities in the wage process.

[13] 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.

[TABLE 3 HERE]

However, although returnees that have primary- and secondary-level educations tend to show higher unobservable traits, the distribution of unobservables is different for college graduates. Figure 3 shows the dissimilarities in the actual and counterfactual distributions at different educational levels in order to better visualize these differences. Figures 3(a) and 3(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 3(c) and 3(d), which show the same distribution for workers that have a secondary-level education. Finally, Figures 3(e) and 3(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 3 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 a 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. Without return migration, the average wage in the population would thus be higher. This is true not only overall, but also looking by education level within the immigrant population. Although returnees are less skilled compared 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. 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 increased competition for immigrant and native low-wage earners.

Return migration influences immigrant inequality. The presented analysis 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 do return to their home countries, too.

The immigrant-native-born 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 of which are shown in all the figures presented earlier).[14] Figure 2 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 3, 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 slightly close 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 3 clearly demonstrates that selection on return migration is inducing middle to top earners to leave the U.S., thus biasing the picture we have cultivated 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 policy might still not fully benefit the U.S., as middle to top wage earners-the most productive workers-would still leave.[15]

7 Robustness Checks

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,

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

[15] 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.

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.[16]

Full Model. The previous discussion constructed the counterfactual and actual distributions based on the estimation of a parsimonious wage equation, where only those variables for which information was given for both stayers and returnees were reported.[17] There may be some concern, however, that a better specified model could change the results. As explained previously, the main problem of using a fully specified wage equation is that no information is present for returnees on the length of stay, the location, and the industry in the U.S. Not to introduce extra uncertainty based on the imputation of these missing variables, I assumed that returnees present similar characteristics to those of average non-returning migrants. Given the previous similarities of the quantiles of $X\hat{\beta}$, this assumption seems reasonable. Panels (a) and (b) in Figure 4 show the actual and counterfactual distribution when more regressors are added into the analysis.[18] All previously drawn conclusions hold for this further specification.

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

Figure 4(c) and 4(d) 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 used to estimate the conditional expectations of the wage process.[19] The parametric results can be seen to be very close to the semiparametric results. Further, a KS test of the equality of the distribution of the parametric and semiparametric models 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 in this context the major conclusions of the paper are also valid. Mexican returnees come from the middle to top part of the distribution, suggesting that in the absence of return migration a larger mass of people would have their wages in the upper part of the wage distribution.

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

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

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

[19] 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. 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.

Illegal Immigration. The problem of the undercount of illegal immigrants is often a concern when using census data, as the fear of deportation 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. By contrast, it seems reasonable to assume that Mexican returnees are well captured in the Mexican census, as the motivation to underreport U.S. experience is not affected by the illegality status in the U.S. Indeed, the undercount of illegal immigrants has been an important influencing factor in Mexican migration research. For example,Fernandez-Huertas Moraga (2011) argues 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.

The first concern is that the observable traits of the U.S. sample used in the present analysis do not capture the characteristics of the Mexican population in the U.S. as a whole. Passel (2006) shows that illegal immigrants tend to be young, low-educated, low-wage workers. If this were the case, however, the results would be biased towards zero, and the differences found herein should be a lower bound for the counterfactual estimates. The second serious concern comes from the non-randomness of the census sample even after controlling for other 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 the 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:

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

Then, the choice of staying in the U.S. is observed only if the individual does appear 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. 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 misestimated. Using again the Law of Total Probability, in fact:

$$P(S_i = 1|Z_i'\alpha) = P(S_i = 1|Z_i'\alpha, C_i = 1)Pr(C_i = 1|Z_i'\alpha) + P(S_i = 1|Z_i'\alpha, C_i = 0)Pr(C_i = 0|Z_i'\alpha),$$

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 one. By doing so, individuals that have large values of $Z_i'\alpha$ were implicitly selected. However, whenever $Z_i'\alpha$ is high, also $W_i'\gamma$ is also high. 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 high, $Pr(C_i = 1|Z_i'\alpha)$ is also high. This implies that in the high probability set the probability of staying is mostly 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 seem to marginalize the problems related to the censoring in the selection rule due to illegal immigration.

As a final check, the analysis was also run excluding the U.S. and Mexican bordering states, where the problem of illegal migration might be more severe. The actual and counterfactual wage distributions are reported in Figures 4(e) and 4(f): 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 undercounted migrants in the U.S. that have worse labor market outcomes.

Circular Migration. The exclusion of the bordering states in panels 5(e) and 5(f) of Figure 4 has the further advantage of reducing concerns that the results are biased by circular migration. Mexican migrants are recognized for engaging in repeated movements into the U.S. (Massey and Espinosa, 1997). Therefore, because the census would be unable to explain whether observed returnees intended to return to the U.S., by excluding the states close to the border, we can exclude those areas in which circular migration is more common. As shown, the results are unchanged in this specification.

Validity of the Assumptions in the Estimation. As a final discussion, it is important to highlight whether the adopted technique's assumptions of exogeneity and heteroskedasticity are reasonable in this context. Endogeneity 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. To check on the validity of this assumption, I compare the estimated unconditional distribution of the error term, $\hat{f}(u^*)$, with the estimated conditional distribution of the error conditioned on the index in the high probability set, $\hat{f}(u^*|Z_i'\hat{\alpha})$. If Z and u^* were dependent these two estimated distributions would differ, and, in particular, the conditional expectation of u^* on $(Z_i'\hat{\alpha})$ would also change at different values of $Z_i'\hat{\alpha}$ also in the high probability set. If Z_i could be treated as independent, the two estimated distributions would still differ slightly because of the inherent randomness of the estimation procedure, but would be relatively close. Figure 5 shows how 'close' $\hat{f}(u^*)$ and $\hat{f}(u^*|Z_i'\hat{\alpha})$ are. 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 as though 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 4 tabulates the D statistic at different deciles of the index. For the majority of the deciles, we cannot reject the null that the conditional and unconditional densities are drawn from the same distributions at the 5% significance level. This comparison hints that selecting on $(Z_i'\hat{\alpha})$ should not be a concern and the exogeneity of the selection rule seems to be verified in the data.

[TABLE 4 HERE]

To further validate this claim, I use as the counterfactual distribution the distribution in the high probability set conditional on the ninth decile of the $(Z_i'\hat{\alpha})$ index, as the KS test found this to be the furthest from the estimated distribution reported in the reminder of the paper. Considering this as the counterfactual of interest, Figure 6 shows the deriving actual and counterfactual wage distributions and their differences. As can be seen, the results remain consistent with the middle to positive selection of returnees.

[FIGURE 6 HERE]

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.

8 Conclusions

The political discussion generated by Mexican migration flows into the U.S. has focused on understanding migration decisions but, until recently, has 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 to the body of knowledge on this topic by analyzing this question through 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 and showed 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 the average earnings ability of immigrants and reduces immigrant wage dispersion. Further, return migration has a greater impact on the tail of the wage distribution. In particular, in the absence of return migration more mass would appear in the upper part of the wage distribution in both very low and very high educational groups, implying up to a 7% increase in the median wages paid to the Mexican migrant population. These results are stable across different wage specifications, samples, and parametric 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 thus 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 about the nature of return migration in the U.S.

9 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
Baseline	0.890	0.901
Age	39.697	0.004***
		(4.29E-04)
Primary	0.068	0.020***
		(0.002)
Secondary	0.647	0.067***
		(0.003)
College	0.282	0.031***
		(0.002)
Married	0.806	-0.006***
		(0.001)
US born spouse	0.677	0.050***
		(0.002)
Child	0.528	-0.063***
		(0.003)
US born child	0.534	0.175***
		(0.004)
N	133,389	133,389

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

Decile	**Actual**	**Counterfactual**	**Log 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	0.000
		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	0.000	0.045

Continue to next page

Continued from previous page			
Decile	**Actual**	**Counterfactual**	**Log Difference**
	Log-Wage		
1	1.618	1.611	-0.007
2	1.853	1.883	0.030
3	2.033	2.090	0.056
4	2.203	2.270	0.067
5	2.360	2.434	0.074
6	2.514	2.585	0.071
7	2.682	2.757	0.075
8	2.896	2.962	0.065
9	3.233	3.274	0.041
Average	2.406	2.451	0.045
	Inequality Measures		
10-90 Wage	1.615	1.663	0.048
10-50 Wage	0.742	0.823	0.081
50-90 Wage	0.873	0.840	-0.033
N	120,205	120,205	120,205

The first column (*Actual*) shows $\hat{Y}$ and $\hat{u}$ for the observed sample. The second column (*Counterfactual*) 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^*.

Table 3: 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.**	**Log Diff**	**Act.**	**Counterfact.**	**Log Diff**	**Act.**	**Counterfact.**	**Log 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	-0.015	-0.674	-0.666	0.007	-0.879	-0.693	0.186
2	-0.488	-0.461	0.026	-0.474	-0.461	0.013	-0.630	-0.475	0.155
3	-0.356	-0.292	0.064	-0.316	-0.292	0.024	-0.419	-0.289	0.130
4	-0.234	-0.147	0.087	-0.183	-0.147	0.036	-0.241	-0.138	0.103
5	-0.119	-0.020	0.099	-0.053	-0.020	0.033	-0.068	0.003	0.071
6	0.007	0.106	0.100	0.076	0.106	0.030	0.104	0.129	0.025
7	0.143	0.249	0.107	0.214	0.249	0.035	0.274	0.271	-0.003
8	0.313	0.418	0.105	0.385	0.418	0.033	0.471	0.428	-0.042
9	0.585	0.657	0.072	0.633	0.657	0.023	0.758	0.656	-0.102
Average	-0.055	0.000	0.055	-0.022	0.000	0.022	-0.053	0.000	0.053
	Log-Wage								
1	1.597	1.574	-0.023	1.682	1.689	0.007	1.793	1.979	0.186
2	1.817	1.838	0.021	1.932	1.945	0.013	2.090	2.245	0.155
3	1.999	2.060	0.061	2.147	2.170	0.023	2.348	2.478	0.130
4	2.153	2.236	0.084	2.313	2.348	0.036	2.568	2.671	0.103
5	2.298	2.386	0.088	2.471	2.504	0.033	2.782	2.852	0.071
6	2.452	2.543	0.091	2.631	2.661	0.030	2.989	3.009	0.020
7	2.612	2.719	0.107	2.794	2.829	0.035	3.180	3.177	-0.003
8	2.811	2.913	0.101	2.992	3.025	0.033	3.402	3.355	-0.046
9	3.104	3.174	0.070	3.269	3.292	0.023	3.703	3.601	-0.102
Average	2.346	2.397	0.051	2.583	2.511	-0.072	2.776	2.828	0.052
	Inequality Measures								
10-90 Wage	1.507	1.600	0.093	1.587	1.603	0.016	1.910	1.622	-0.288
10-50 Wage	0.701	0.812	0.111	0.789	0.815	0.026	0.989	0.873	-0.115
50-90 Wage	0.806	0.788	-0.018	0.798	0.788	-0.010	0.921	0.748	-0.173
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^*.

Table 4: Kolmogorov-Smirnov Test for the difference in $f(u^*|Z_i'\hat{\alpha}_{[q]})$ and $f(u^*)$.

	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***

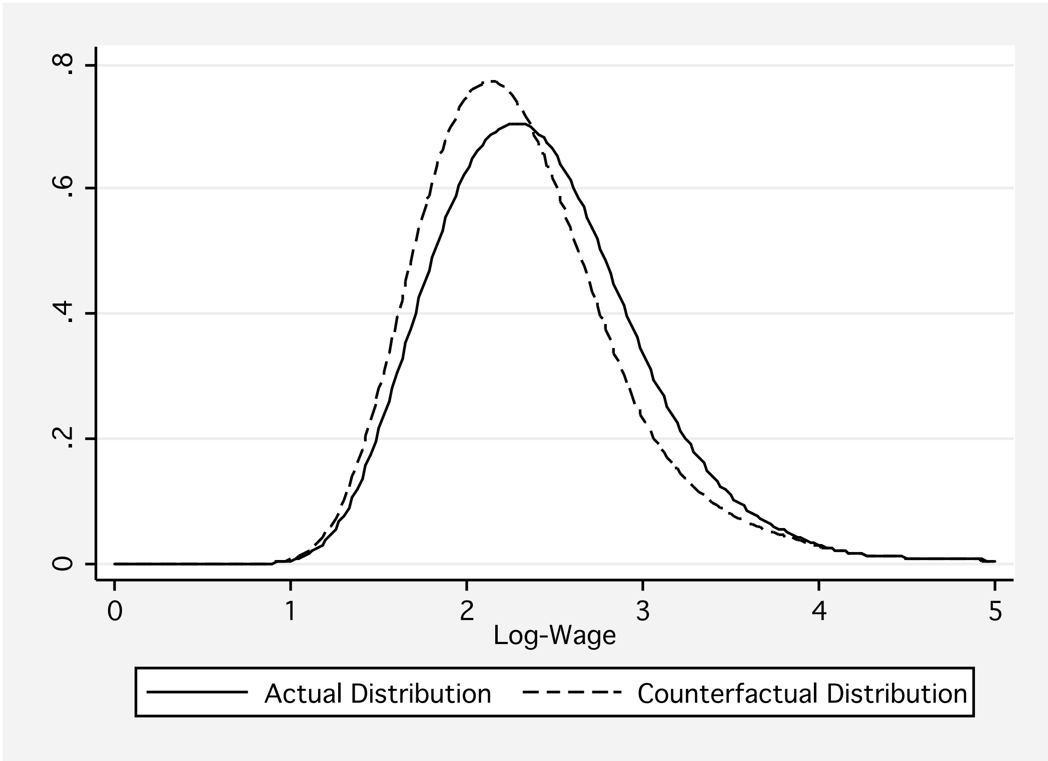
Significance levels: *: 10%, **: 5%, ***: 1%.

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

The test was constructed comparing the distribution of u^* ($f(u^*)$) with the distribution of u^* for individuals in each decile $[q]$ of the $(Z_i'\alpha)$-index ($f(u^*|Z_i'\hat{\alpha}_{[q]})$) . If $Z_i'\hat{\alpha}$ was exogenous, these two vectors should be drawn from the same distribution.

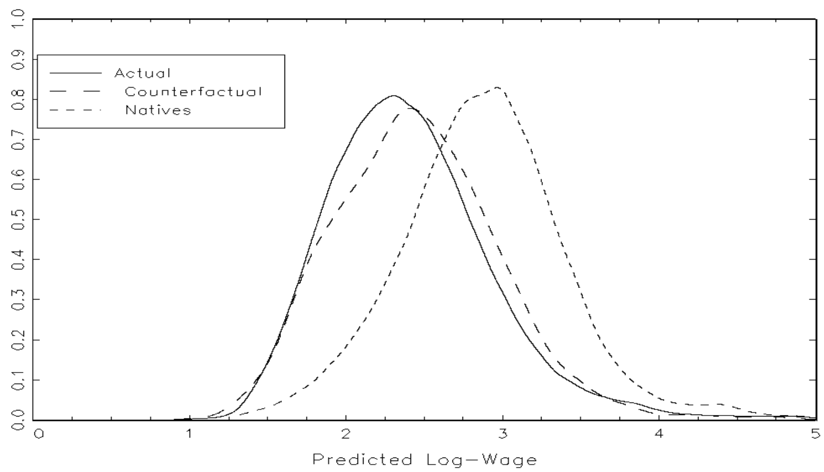
10 Figures

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

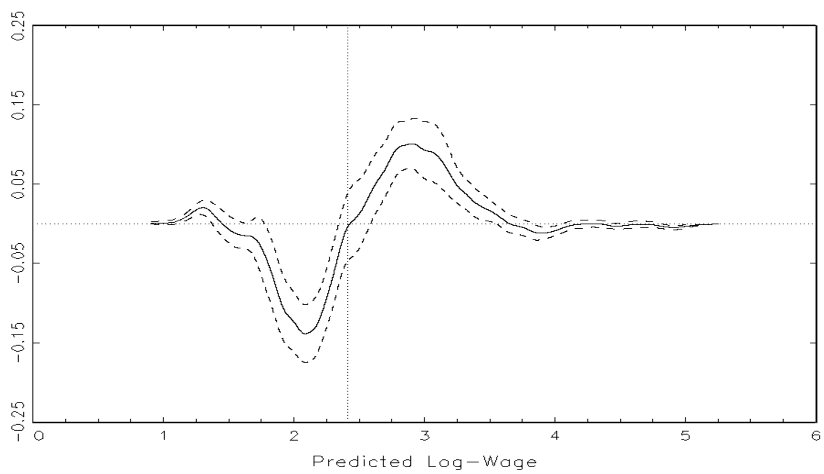


The actual distribution represents the distribution of wages for Mexican-born workers in the U.S. The counterfactual distribution represents the distribution of wages that would have occurred in the U.S. if Mexican-born workers had the characteristics of the returnees. These distributions are obtained 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 Counterfactual Log-Wage Distributions (a) and their Difference (b), Parsimonious Model, Men, 35-55 Years Old.



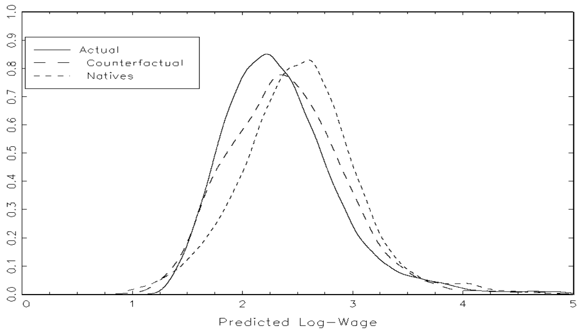
(a) Actual (Solid Line) and Counterfactual (Dashed Line) Wage Distribution



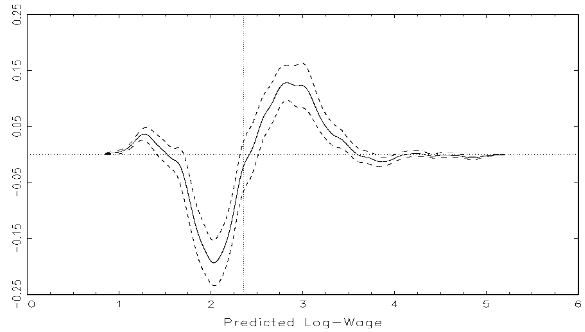
(b) Difference in Counterfactual and Actual Distributions

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 2 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.

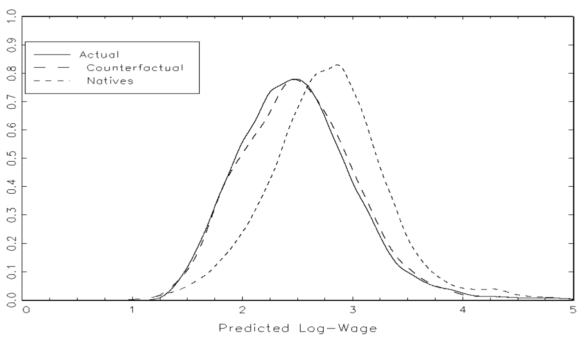
Figure 3: Estimated Actual and Counterfactual Log-Wage Densities for Mexican immigrants with Primary (a), Secondary (c), and College Education (e), Parsimonious Model, Men 35-55 Years Old.



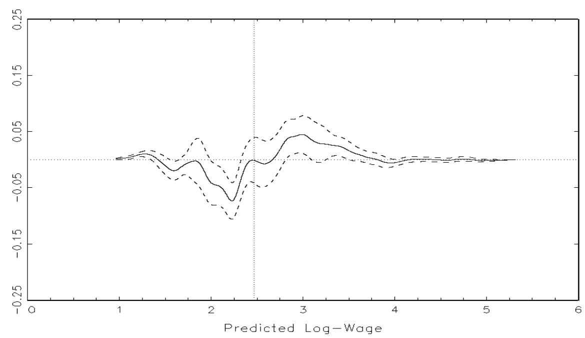
(a) Primary Education, $n = 27,403$



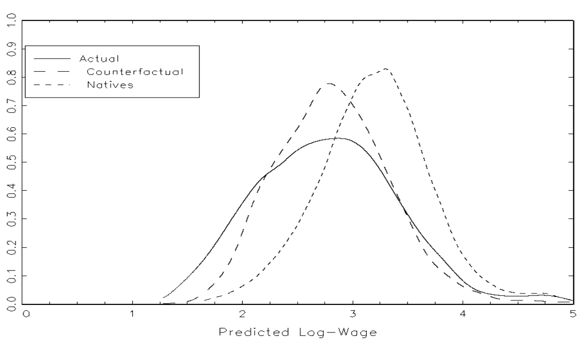
(b) Difference in Counterfactual and Actual Distributions, Primary Education



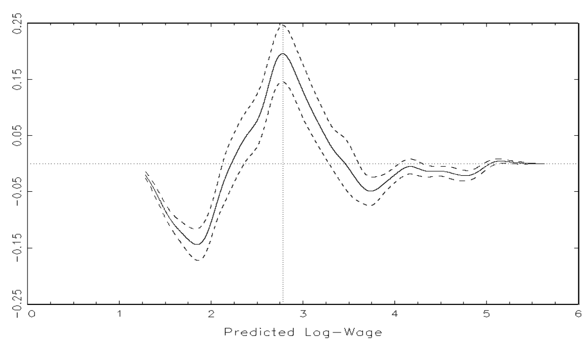
(c) Secondary Education, $n = 18,811$



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



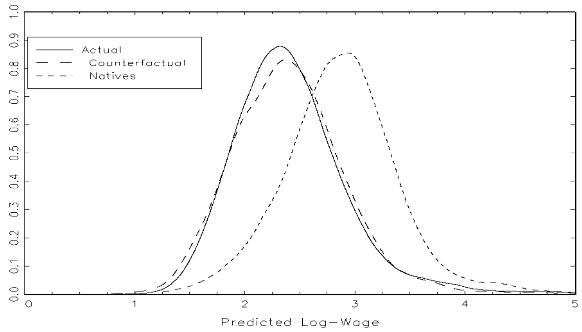
(e) College Education, $n = 3,001$



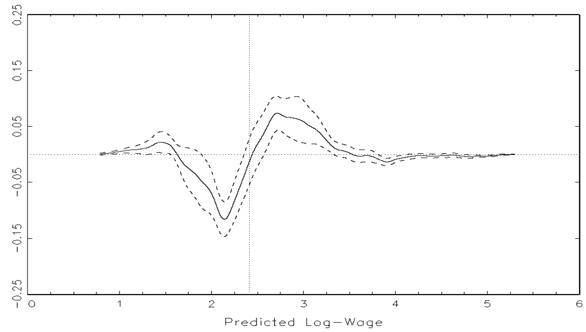
(f) Difference in Counterfactual and Actual Distributions, College Education

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. 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).
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.

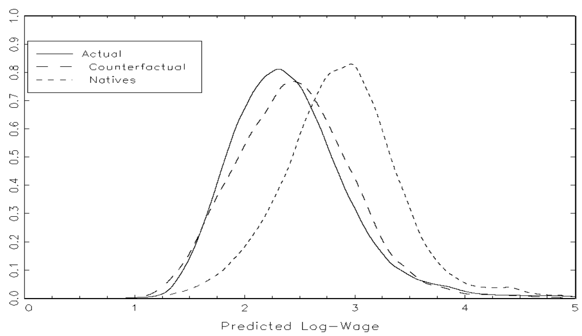
Figure 4: Counterfactual and Actual Log-Wage Densities and their Differences for Mexican immigrants, Full Model (a)-(b), Parametric Model (c)-(d), and Model Excluding Bordering States (e)-(f).



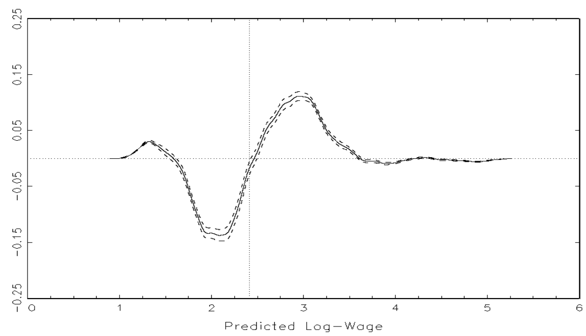
(a) Full Model



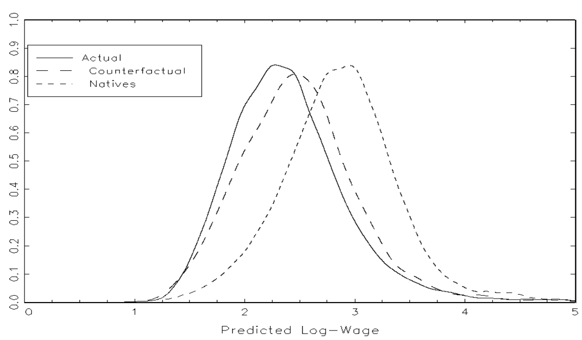
(b) Full Model



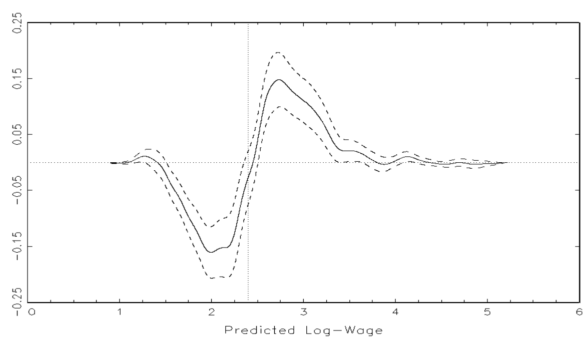
(c) Parametric Model



(d) Parametric Model



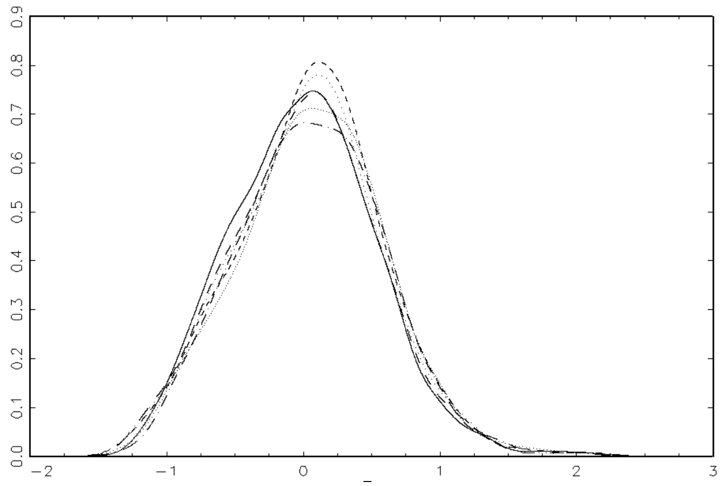
(e) Model excluding Mexican-U.S. Bordering States



(f) Model excluding Mexican-U.S. Bordering States

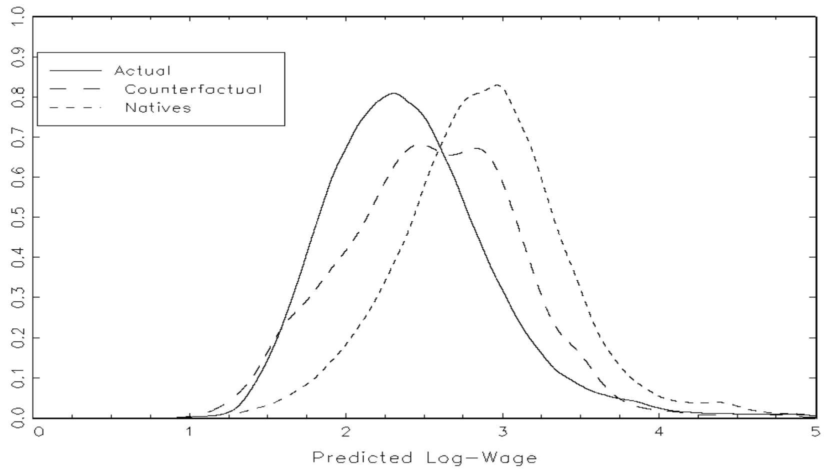
The Actual distribution represents the distribution of wages for Mexican-born workers currently residing in the U.S. The Counterfactual distribution represents the distribution of wages in the U.S. for Mexican-born workers if all migrants settled permanently, i.e., if no return migration occurred between 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 (c)-(d) estimates the distribution using parametric techniques. Table C.3 reports the regression results of this estimation technique. Model (e)-(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)-(b): 3.10, in model (c)-(d): 5.43, in model (e)-(f): 4.89. Critical value at 1%: 1.63.

Figure 5: Counterfactual Distributions conditioning on different deciles of the $Z_i'\hat{\alpha}$ index in the High Probability Set, $f(u^*|Z_i'\hat{\alpha}_{[q]})$

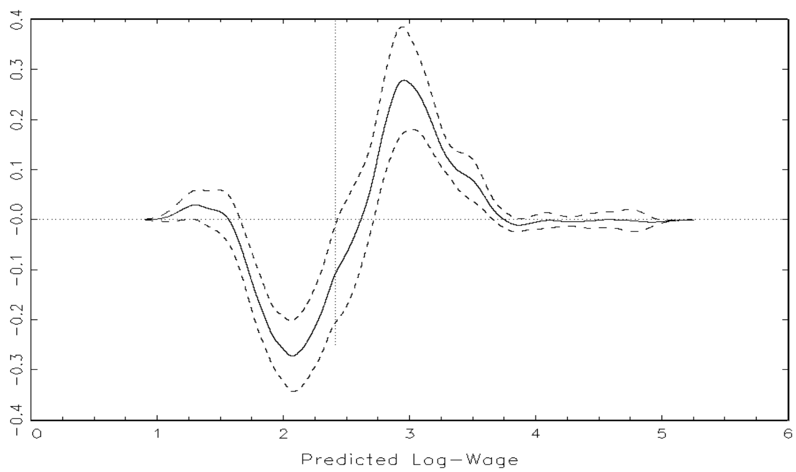


The figure compares the distribution of u^* for individuals in various deciles $[q]$ of the $(Z_i'\alpha)$-index $(f(u^*|Z_i'\hat{\alpha}_{[q]}))$. If $Z_i'\hat{\alpha}$ was exogenous, these conditional distributions should be close.
Table 4 reports the Kolmogorov-Smirnov test for equality of the $f(u^*)$ and $f(u^*|Z_i'\hat{\alpha})$ distribution functions.

Figure 6: 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. In other words, the figure compares $f(u^*|Z_i'\hat{\alpha}_{[9]})$ with $f(u^*|S=1)$. 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.

References

Ambrosini, J. William and Giovanni Peri, "The Determinants and the Selection of Mexico-US Migrations," *The World Economy*, February 2012, *35* (2), 111–151.

_ **, Karin Mayr, Giovanni Peri, and Dragos Radu**, "The Selection of Migrants and Returnees: Evidence from Romania and Implications," Working Paper 16912, National Bureau of Economic Research 2011.

Andrews, Donald W. K. and Marcia M.A. Schafgans, "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 1998, *65* (3), 497–517.

Angelucci, Manuela, "U.S. Border Enforcement and the Net Flow of Mexican Illegal Migration," *Economic Development and Cultural Change*, 2012, *60* (2), 311 – 357.

Borjas, G. J., "Self-selection and the Earnings of Immigrants," *American Economic Review*, September 1987, *77*, 531–553.

Borjas, George J., "The Relationship between Wages and Weekly Hours of Work: the Role of Division Bias," *The Journal of Human Resources*, 1980, *15* (3), 409–423.

_ **and Bernt Bratsberg**, "Who Leaves? The Outmigration of the Foreign-Born," *The Review of Economics and Statistics*, February 1996, *78* (1), 165–176.

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

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

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 1996, *64* (5), 1001–1044.

Heckman, James J., "Varieties of Selection Bias," *The American Economic Review*, 1990, *80* (2), 313–138, Papers and Proceedings.

Hu, Wei-Yin, "Immigrant earnings assimilation: estimates from Longitudinal data," *The American Economic Review*, 2000, *90* (2), 368–372.

Ibarraran, Pablo and Darren Lubotsky, "Mexican Immigration and Self-Selection: New Evidence from the 2000 Mexican Census," in "Mexican Immigration to the United States" NBER Chapters, National Bureau of Economic Research, Inc, 2007, pp. 159–192.

Kaestner, Robert and Ofer Malamud, "Self-Selection and International Migration: New Evidence from Mexico," Working Paper 15765, National Bureau of Economic Research February 2010.

Klein, Roger W. and Richard H. Spady, "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, March 1993, *61* (2), 387–421.

_ , Chan Shen, and Francis Vella, "Semiparametric Selection Models with Binary Outcomes," Discussion Paper 6008, Institute for the Study of Labor (IZA) 2011.

Lacuesta, Aitor, "A Revision of the Self-Selection of Migrants Using Returning Migrants Earnings," *Annals of Economics and Statistics*, 2010, *97/98*, 235–259.

Lindstrom, D. P. and D. S. Massey, "Selective Emigration, Cohort Quality, and Models of Immigrant Assimilation," *Social Science Research*, 1994, *23* (4), 315 – 349.

Lubotsky, Darren, "Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings," *Journal of Political Economy*, 2007, *115* (5), 820–867.

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

McKenzie, David and Hillel Rapoport, "Self-selection Patterns in Mexico-U.S. Migration: the Role of Migration Networks," *The Review of Economics and Statistics*, 2010, *92* (4), 811–821.

Moraga, Jesus Fernandez-Huertas, "New Evidence on Emigrant Selection," *The Review of Economics and Statistics*, 2011, *93* (1), 72–96.

Passel, Jeffrey S., "The Size and Characteristics of the Unauthorized Migrant Population in the U.S," Research Report, Pew Hispanic Center 2006.

Reinhold, Steffen and Kevin Thom, "Temporary Migration, Skill Upgrading, and Legal Status: Evidence from Mexican Migrants," *MEA DP*, 2009, *182.*

Robinson, P.M., "Root-N Consistent Semiparametric Regression," *Econometrica*, July 1988, *56* (4), 931–954.

Schafgans, Marcia M. A. and Victoria Zinde-Walsh, "On the Intercept Estimation in the Sample Selection Model," *Econometric Theory*, February 2002, *18* (1), 40–50.

Silverman, Bernard W, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

Thom, Kevin, "Repeated Circular Migration: Theory and Evidence from Undocumented Migrants," August 2010. Working Paper.

Yang, Dean, "Why Do Migrants Return to Poor Countries? Evidence from Philippine Migrants' Responses to Exchange Rate Shocks," *Review of Economic and Statistics*, November 2006, *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:

$$S_i = \begin{cases} 1 & 1 + X_1 + 2X_2 \geq \epsilon \\ 0 & 1 + X_1 + 2X_2 < \epsilon \end{cases}$$

$$Y_i = 1 + X_1 + u_i \qquad \text{if} \quad S_i = 1$$

Here X_1, X_2, u and ϵ are standard normal random variables. For each iteration in the Monte Carlo experiment, I calculate the deciles of the distribution of u^*, estimated as explained above, 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.1 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.

Table A.1: 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:

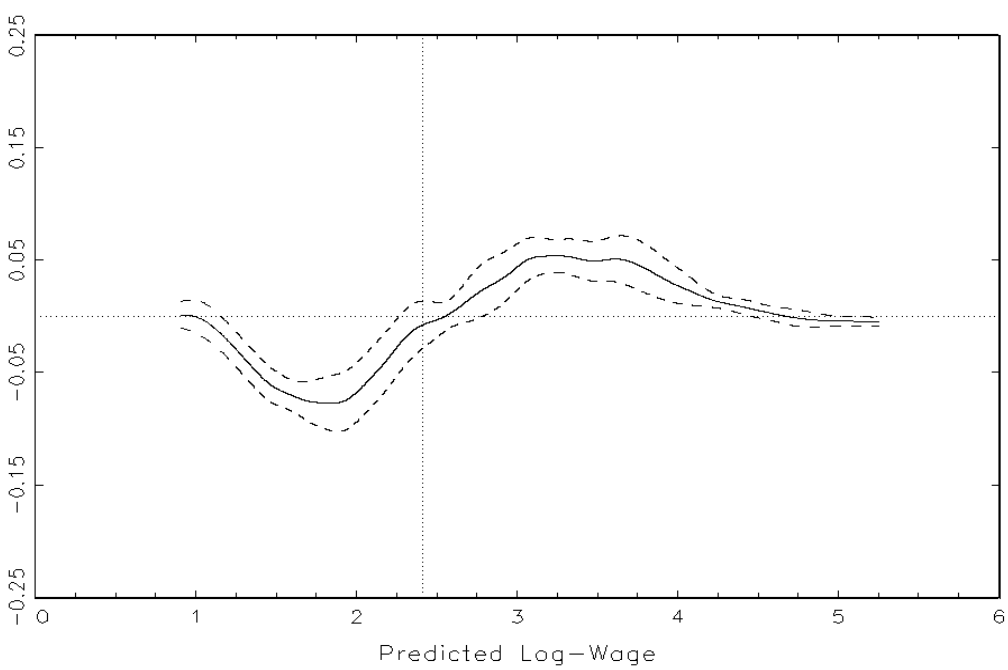
$$Y_i = X_i'\beta_0 + c_0 + G(Z_i'\alpha_0) + u^* k(X\delta_0),$$

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 then recovers the distribution of u^*.

The figure below shows the results. As it can be seen the main conclusions of the paper are still obtained.

Figure 7: Difference in Counterfactual and Actual Distributions accounting for Heteroskedasticity



Standard errors have been bootstrapped (100 replications).

C Additional Tables

Table C.2: Demographic and socio-economic characteristics, Native Born and Foreign Born Men, 25-55 Years Old

Variable	Natives	All Mexican Born	Stayers	Returnees
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)	-
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***

Continue to next page

Continued from previous page				
Variable	**Natives**	**All Mexican Born**	**Stayers**	**Returnees**
	(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)
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)
Spouse US	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	0.004***	0.022***
	(0.001)	(0.002)
Age Sq	-	-2.6E-04***
	-	(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)
Spouse US born	0.051***	0.057***
	(0.001)	(0.005)
Child	-0.073***	0.100***
	(0.001)	(0.004)
Child US born	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.