J Popul Econ (2014) 27:923–960
DOI 10.1007/s00148-014-0524-x

ORIGINAL PAPER

Empirical characteristics of legal and illegal immigrants in the USA

Vincenzo Caponi · Miana Plesca

Received: 4 April 2013 / Accepted: 27 May 2014 /
Published online: 20 June 2014

Abstract We combine the New Immigrant Survey (NIS), which contains information on US legal immigrants, with the American Community Survey (ACS), which contains information on legal and illegal immigrants to the USA. Using an econometric methodology proposed by Lancaster and Imbens (J Econ 71:145–160, 1996) we compute the probability for each observation in the ACS data to refer to an illegal immigrant, conditional on observed characteristics. These results are novel, since no other work has quantified the characteristics of illegal immigrants from a random sample representative of the population. Using these conditional probability weights on the ACS data, we are able to uncover some interesting facts on illegal immigrants. We find that, while illegal immigrants suffer a large wage penalty compared to legal immigrants at all education levels, the penalty decreases with education. We also find that the total fertility rate among illegal immigrant women is significantly higher than that among legal ones, in particular for middle and higher educated women. Looking

Responsible editor: Klaus F. Zimmermann

V. Caponi (✉)
CREST (Ensai), Rennes, France
e-mail: vincenzo.caponi@ensai.fr

V. Caponi
Ryerson University, Toronto, Canada

V. Caponi
Kiel Institute for the World Economy, Kiel, Germany

V. Caponi
IZA, Bonn, Germany

M. Plesca
Department of Economics, University of Guelph, Guelph, Canada
e-mail: miplesca@uoguelph.ca

at the sector of activity, we document that the sectors attracting most illegal immigrants are constructions and agriculture. We also generate empirical distributions for state of residence, country of origin, age, sex, and number of legal and illegal immigrants. Our forecasts for the aggregate distribution of legal and illegal characteristics match imputations by the Department of Homeland Security.

Keywords Legal immigrants · Illegal immigrants · Contaminated controls

JEL Classifications J15 · F22

1 Introduction

In the recent history of immigration, a pressing concern has been the increasing trend of illegal immigration, especially in developed countries. Most countries have in place immigration policies designed to welcome immigrants under terms deemed beneficial for the host country. Illegal immigration is often seen as problematic because, by its nature, it circumvents the control of policy makers. Immigrants eluding legal immigration channels are those who would not be accepted otherwise, because they are either in excess number or of different quality than desired by the destination country. Yet, every day, thousands of persons cross the border to start their new lives as illegal immigrants. In this paper, we provide answers to some questions about the illegal immigrants: how many are they, what are their characteristics, how do they differ from the legal immigrants, what determines their human capital, and how is their human capital rewarded in the labor market.

We provide and apply a methodology able to separate the legal from the illegal immigrants in a large US national survey. Using information on immigrants in the USA from the American Community Survey (ACS) and the New Immigrant Survey (NIS), we are able to identify a set of conditional probability weights determining whether individuals are legal or illegal immigrants based on their observed characteristics. Since the ACS refers to the entire population of immigrants, legal and illegal, and the NIS refers to the subpopulation of legal immigrants only, a difference between the two datasets can provide information on the characteristics and number of illegal immigrants. Once we exploit the differences in observed characteristics between legal and illegal immigrants, we can use these conditional probability weights to compute statistics for the legal and illegal immigrants. Among other statistics, we generate the empirical distributions of educational attainment, fertility, occupations, and wages separately for legal and illegal immigrants. This ensures that we get a better and objective understanding of who the illegal immigrants are and how they compare with the legal ones.

There is a large related literature on the economic outcomes of immigrants to the USA, and in particular, of Mexican immigrants. Most of this literature uses the Current Population Survey (CPS) or Census data to investigate economic outcomes for immigrants, sometimes in relation to those of natives, without distinguishing between legal and illegal immigration status. Our paper is complementary to that literature because we focus primarily on a methodology to identify the illegal and legal immigrants in a large dataset comparable with the US Census.

A different approach has been to use information on legal and illegal migrants from Mexico using data available from the Mexican Migration Project (MMP), whose aim is to collect information on the legal and illegal Mexican migration to the USA. The main drawback of the MMP is that it is not a random survey; instead, it selects certain rural communities from Mexico where the population is more likely to emigrate. The MMP interviews households from these prespecified communities with high out-migration propensity and asks questions about Mexican residents living in Mexico and in the USA. [1] Because of the nonrandom design, any research conclusions using the MMP data cannot be generalized over all Mexican immigrants to the USA, and even less so over the entire population of immigrants to the USA.

Despite the fact that the MMP sample is not a representative random survey, it still provides some relevant information about the migration histories of the respondents, including whether they arrived in the USA legally or illegally, and their socioeconomic characteristics. From research using the MMP, we have learned that most of the Mexican illegal migration is return migration, with about 85 % of illegal immigrants returning home to Mexico; that the majority of illegal Mexican migrants are young males working in farming; that the recent trends see a shift away from farming and construction into services; that women have very different migration patterns than men, in the sense that they are more likely to follow a spouse to the USA and are more likely to stay in the USA once they arrive; and that the traditional rural sources of Mexican immigration are shifting toward urban areas (Durand and Massey 2006).

In our analysis, we quantify some of these patterns using the entire immigrant population in the USA, with Mexican immigrants as a subpopulation of interest for whom we allow heterogeneous effects. We document similar broad facts for the Mexican immigrants as for the overall immigrant population: illegal immigrants are more likely to be male, single, or married but with no spouse present in the USA, and earning less than legal immigrants. We find some differences in the propensity to immigrate illegally between Mexican and other immigrants by education. Low education levels predict a higher probability of being illegal for immigrants in general, but the opposite pattern is true for Mexican immigrants, for whom being educated increases the likelihood of immigrating illegally to the USA.

The other literature that provides estimates of the undocumented foreign-born population in the USA comes mostly from demographers who rely on the residual method. This method compares data from a representative survey, most often the US Census or the CPS, with aggregate statistics on legal entrants provided by the Department of Homeland Security (DHS). The population survey is giving information on the size and the characteristics of the total population of foreign-born residents in the USA at a given point in time. The DHS provides aggregate statistics for the inflows and outflows of individuals who are legally entitled to reside in the USA. The aggregate measure of the unauthorized migrant population is given by the total foreign

[1] Since its inception in 1987, the MMP has surveyed every year between four to eight communities during earlier survey years and between two to five communities during more recent survey years, to an overall total of 81 selected communities. For each household head and spouse, full migration and labor market histories are constructed from recall information; other household members are also interviewed about their first and their last trip to the USA.

population minus the sum of all current and previous net flows of legal immigrants, accounting for attrition through mortality. This measure should give a reasonably accurate count of the size of the illegal immigrant population (Passel 2006; Passel et al. 2004). Nevertheless, because of the aggregate nature of the DHS statistics, the residual method can only give the number of legal and illegal immigrants across broad aggregate dimensions, while important socioeconomic characteristics such as age, education, or marital status remain missing from the analysis.[2]

Our proposed approach is very similar in spirit to the residual method, in that it compares two sets of information: one on legal immigrants (the NIS) and the other on the general population of foreign-born individuals (the ACS).[3]

The methodology we implement here was proposed by Lancaster and Imbens (1996) to deal with applications when the treated population can be identified from a sample of treated observations, or "cases," while the "control" population cannot be immediately identified; instead, a mixed sample of case and control individuals is observed. A random observation is drawn either from the case sample or from the mixed case-control sample, based on a Bernoulli process. Given covariates X, a likelihood function is written. Lancaster and Imbens (1996) provide moment conditions which are equivalent to maximizing the likelihood function. From the moment conditions, we can generate a set of case and control probabilities conditional on observed characteristics X. In applying their methodology to our problem, the "cases" are legal immigrants from the NIS survey of green card holders, whom we observe free of contamination. The mixed "case-control" observations are the immigrants from the ACS data: since we cannot identify ex ante who are legal immigrants—cases—and who are illegal immigrants—"controls"—in ACS, we have a mixed case-control sample.

Following the GMM methodology by Lancaster and Imbens (1996), we generate the conditional probability for each observation to be a legal immigrant case. We can use these probability weights to compute all sorts of statistics separately for legal and illegal immigrants, from any data with information of all immigrants, such as the ACS, Census, CPS, or some other representative dataset. Any other researcher can also generate these conditional weights using the probit coefficients we report here and immigrant characteristics from a representative dataset of their own choice.

[2] Related literature investigates the effect of US border enforcement in stemming the flows of illegal Mexican immigration. This approach, referred to as the Apprehensions Method, does not provide an estimate of the number of illegal immigrants in the USA, nor of their characteristics. However, it does provide information on the change in time of the inflow of legal and illegal immigrants (Rosenblum 2012).

[3] Somewhat related to our approach, Burtless and Singer (2011) combine data from the MMP with CPS data to get a measure of how many illegal Mexicans contribute to Social Security (being illegal, they have no hope of withdrawing benefits, despite contributing to Social Security). Because they need to identify who in the CPS data is an undocumented immigrant, they use a matching algorithm which they call "cold decking" to infer who in the representative CPS data would be a legal or an illegal Mexican migrant based on the observed characteristics of legal and illegal migrants in the MMP data. While their approach has a very different context, it still suffers from the fact that the MMP is a nonrandom sample and the characteristics of legal and illegal migrants in MMS may be different from the overall characteristics of legal and illegal migrants in the Mexican migrant population.

Our main contribution to the literature on legal and illegal immigrant characteristics coming from demography studies (e.g., Passel (2006)) resides in the versatility of the conditional probability weights, which apply to each individual observation. We use representative *microdata*, which allows us not only to estimate aggregates, such as the number of immigrants residing illegally, but also their personal characteristics, labor market performance, and human capital determinants, along with any other information available in both surveys. To this extent, we document some interesting facts on illegal immigrants. For example, we find that while illegal immigrants suffer a large wage penalty compared to legal immigrants at all education levels, the penalty decreases with education. We also find that the total fertility rate among illegal immigrant women is significantly higher than that among legal ones, and that is particularly true for middle and higher educated women. Finally, we look at the sector of activity, and we find that the constructions sector is the sector that most attracts illegal immigrants and that most of the immigrants in agriculture are also illegal.

Relative to the other strand of immigration literature coming from studies using the MMP, our contribution is to use *representative random samples* of immigrants. Consequently, our results hold generally for all immigrants to the USA, and not only for a subsample of Mexican immigrants selected in a particular way.

The paper proceeds as follows. Section 2 describes the two datasets used in the analysis, ACS and NIS, and Section 3 describes how we adapt the contaminated-controls methodology proposed in Lancaster and Imbens (1996) to identify the propensity to be an illegal immigrant. Section 4 presents our main results, and Section 5 provides some sensitivity checks and examples of the further analysis that can be pursued given our identification of the legal/illegal conditional probabilities. Appendix 1, Appendix 2, Appendix 3, and Appendix 4 provide more sensitivity checks. Section 7 concludes.

2 Data and sample statistics

We describe here the two datasets we base our analysis on. The NIS samples legal permanent residents (LPR) in the USA who acquired legal status, or green cards, in 2003. They are our sample of legal immigrants in the USA. We compare them with a sample of all immigrants in the USA, either legal or illegal, who are surveyed yearly in the ACS. The difference in these two populations, all immigrants vs. legal immigrants, can give us a measure of the characteristics of legal and illegal immigrants to the USA. In this section, we detail the data filtering performed to ensure the two samples are comparable and representative of their underlying populations.

2.1 NIS

After a pilot project in 2001, the NIS started officially with its first wave in 2003. Within this flow of new legal residents, some individuals were already temporary residents of the USA, while others entered the USA for their first time only after having received their green card. Table 1 reproduces the unweighted and weighted frequencies for each class of immigration as tabulated by the NIS.

Table 1 NIS class of admission—adult sample

Visa types in data	Unweighted (%)	Weighted[a] (%)
Spouse of US citizen	16.7	34.2
Spouse of legal permanent resident	2.4	2.4
Parent of US citizen	11.6	11.9
Child of US citizen	3.3	3.4
Family fourth preference	6.2	6.4
Employment preferences	19.5	9.6
Diversity immigrants	16.9	8.1
Refugee	6.5	6.6
Legalization	7.7	8.0
Other	9.2	9.4
Total	100	100

[a] Uses NIS survey weights

Most of those who obtained their green card while already temporary residents qualified for the permanent status under the class of family reunification. The NIS under-samples the family reunification, and in particular those admitted to permanent residence because of marriage to US citizens—who are also more likely to have already been residing in the USA under legal visas. In contrast, people admitted through the visa lottery (Diversity Immigrants) and on the basis of arranged employment are over-sampled. To keep our experiment as clean as possible, we restrict our attention to the flow of new legal immigrants who *entered* the USA in 2003, eliminating those who had been legally residing in the USA and changed their visa status to green card. This allows us to compare them with the overall flow of all new immigrants in the USA in 2003.

2.2 ACS

The ACS samples households randomly across the entire population of US residents, without distinction between citizens or aliens. For the foreign-born population, it makes no distinction between legal or illegal status.[4] Because of language barriers, immigrants are twice more likely to fail to complete the mail-in questionnaires. Consequently, they are more likely to have their data collected during a second, in-person

[4] The ACS, which has been piloted since 1996, is intended as a replacement for the Census long form. While estimates from ACS are slightly less precise than those from the Census long form, a comparison of data from the 2000 Census with data from the 1999–2001 ACS indicated that data quality from ACS was very close to the one in the Census (Camarota and Jeffrey 2004). The obvious advantage of ACS over Census data is that it is a yearly survey, thus providing in a timely manner information on the characteristics of the foreign-born population. Compared to the CPS, estimates from ACS on the characteristics of the immigrant population are more precise.

phase of the interviewing process, resulting in data of better quality (albeit at the cost of higher standard errors, because only about a third are selected for in-person interviews).

In the ACS, we restrict our attention to foreign-born individuals who immigrated in the USA in 2003. Since our final analysis includes both data surveys, ACS and NIS, we need to be sure that the likelihood that each observation is drawn from the population is the same except for the legal/illegal status. That is, if we knew who was legal in the ACS data, we would like to make sure that this observation had exactly the same probability to be observed in the ACS as in the NIS.

Another issue related to weights refers to the possibility that nonresponse rates differ between legal and illegal immigrants. While we know little about the overall response rate of legal and illegal immigrants in the ACS survey, we do have some information about the response rates in the 2000 Census on which the ACS is based. The Department of Homeland Security reports aggregate estimates of the overall documented and undocumented immigrant population which correct for a nonresponse rate in the ACS of about 2.5 % for documented and 10 % for undocumented immigrants.[5] To account for the differential nonresponse rates, we need to modify the survey weights to better reflect the undercounting of illegal immigrants relative to legals in the ACS. We deal with the theoretical implications of weighting for different nonresponse rates in Section 3.1 where we show how to modify the survey weights such that they account for differences in response rates. We present in parallel results using two different sets of weights—one corrected for differences in nonresponse rates, the other uncorrected.

A final weight issue refers to how the survey weights are represented in the two datasets. The NIS uses weights that represent the inverse of the probability for each observation to be randomly chosen, normalized such that the weights sum to the sample size. The ACS has a similar weighing scheme, except that the normalization is done such that the sum of all the weights reproduces the overall population in the USA. We make all weights consistent across the surveys by normalizing the ACS weights, at the same time accounting for differences in the sampling frame.

2.3 Filtering the data: visa holders

One major concern in our analysis is the fact that the ACS includes not only legal and illegal immigrants, but also two other types of foreign-born residents in the USA indistinguishable from the legal and illegal types: refugees and temporary visa holders. There are many different types of temporary visitors that can enter with or without a visa in the USA. According to the DHS yearbook of immigration statistics, of the 27,849,443 nonimmigrants who entered the USA in 2003, the vast majority,

[5] See Hoefer et al. (2008).

24,913,182 were temporary visitors for pleasure or business reasons (or transiting to other locations).[6]

Tourists are unlikely to be counted in the Census or the ACS, and therefore we do not worry about them. Students, mostly under F and J visas, are another big component of the total nonimmigrant population. The DHS reports that in 2003, 946,577 individuals entered as students or exchange visitors.[7] Students can be easily identified in the ACS and the NIS and excluded from the analysis.[8]

The remaining visa holders are more heterogeneous. If we exclude temporary residents belonging to military personnel (NATO visas) and foreign government personnel (A visas), we are left with about one million visa holders. About 10 % of this group(120,000) is represented by H-2 visa holders, who are mostly workers in the agricultural sector (15,000) and in other services (105,000). This type of visas, like the exchange working visas, is given for only 1 year and only exceptionally can be renewed for a maximum stay of 3 years. The remaining nonimmigrants are in occupations that require high skills (850,000 temporary visa holders).[9]

In our analysis, we cannot distinguish who in the ACS belongs to the temporary visa holder category. In order to minimize the extent of contamination, we exclude students from the analysis. Moreover, to identify the characteristics of legal and illegal immigrants, we use data from the ACS 2007 (or, with similar results, from ACS 2006) restricted to foreign-born individuals who have immigrated in 2003. This should exclude those temporary workers in low-skill occupation with visas that are unlike to last for more than a year (J-1 and H-2). In fact, if there are any such individuals still residing in the USA, they likely have overstayed their visa term and become illegal immigrants.

However, visas can also be renewed. Most often a renewal occurs when there is some continuity in the occupation taken by the temporary resident, and often the temporary resident changes the status to a permanent one after one or few renewals.

[6] We are less concerned about refugees because they are fewer than temporary visa holders. According to the DHS (Yearbook of Immigration Statistics), there were 40,705 refugee status applications in 2003, out of which 25,329 were approved. Therefore, we expect our sample to have about 25,000 nonpermanent resident immigrants who entered the USA in 2003 and who live in the USA legally as refugees, representing less than 1 % of the immigrants entered in 2003. The other 15,000 applicants who were rejected may also still live in the USA in 2007; however, they are more likely to be undocumented.

[7] Exchange visitors are for most part students under J-1 visas. A small fraction of them can be workers, who typically are in short work programs to obtain a J-1 visa. These programs take no longer than 2 years to complete and often are only summer jobs.

[8] We conduct further sensitivity analysis to look at the effect of alternative assumptions on the effective legal status of students. That is, we compare our aggregate estimates with the DHS under three hypotheses: (i) students are effectively all legally residing in the USA; (ii) students are all "posers" and effectively illegal immigrants (this happens when students have legal visas but they do not limit their activity to what the visa prescribe, i.e., they work instead than attend a school program), and (iii) students are similar to the rest of the foreign-born population, in which case we allow our methodology to assign probabilities of being legal based on their characteristics. We show that our estimates are closest to the DHS when we assume that students are all residing legally in the USA.

[9] Out of them, 434,281 nonimmigrants are intracompany transferees and their spouses (L visas); about 30,000 representatives and staff of international organizations; 10,000 persons with extraordinary ability in the sciences, arts, education, business or athletics (O visas); 12,000 representatives of foreign media (I visas); and 361,470 workers of distinguished merit and ability (H-1 visas).

This, however, is most likely to happen for highly skilled workers in high occupations (H1 and L visas), the group that, along with students, is the largest among nonimmigrants. As a further measure trying to exclude temporary visa holders from our analysis, in sensitivity analysis, we exclude all persons from high-skill occupations from the ACS and the NIS samples. Once we do that, our results change only slightly, and the estimated fraction of legal immigrants moves a little closer to the DHS benchmark.

Consequently, in the main analysis, we use ACS 2007 compared to NIS to obtain the probabilities, conditional on observable characteristics, of each observation to belong to a legal or illegal immigrant. We use these probabilities as weights in 3 years of ACS data, 2005 to 2007, to predict the distribution of legal and illegal immigrant characteristics in the immigrant population and to compare our results with some known statistics published by the DHS.

2.4 Other data manipulation

While the two datasets, NIS and ACS, are very comparable, we need to make sure that all variable definitions are consistent across the two sets.

In ACS, we construct the hourly wage variable by dividing the total yearly labor income by total hours worked in the year (hours per week times weeks a year). Note that yearly labor income will have more observations than hourly wages because some individuals report positive labor earnings but zero weeks or hours. In NIS, wages/salaries are reported either as hourly wage, or as salaries which have attached a salary schedule; if the latter, we convert earnings into the hourly wage measure.

In both datasets, we aggregate the information on education into the following categories: less than 4 years (“none”); 5 to 8 years (“elementary”); 9 to 12 years, no diploma (“junior high”); high school diploma; college (postsecondary education up to Bachelor’s); and higher education (postgraduate, including Master’s, Ph.D., and professional degrees).

The ACS distinguishes between married individuals whose spouse is present or not. We reconstruct this information in NIS as well, so now the marital status categories are “Married spouse present,” “Married spouse absent,” “Widowed,” “Divorced,” “Separated,” “Never married.” In the analysis, we aggregate some of these categories (such as divorced, widowed, or separated) into one (Table 2).

3 Methodology

The methodology we apply is largely based on that of Lancaster and Imbens (1996) and Ridder and Moffitt (2007). Let s be a “stratum” indicator that takes a value equal to 1 when an observation belongs to the NIS dataset and 0 when it belongs to the ACS. Let Y be a random variable for immigrant status which takes values 1 for legal immigrant and 0 for illegal; then, we know that if $s = 1$, then $Y = 1$ as well. However, when $s = 0$, we do not know what Y is since both legal and illegal

Table 2 Summary statistics from ACS 2007 and NIS

	Including professionals		Excluding professionals	
	ACS	NIS	ACS	NIS
Gender (female)	0.490	0.554	0.374	0.397
Age	31.915	41.754	30.491	36.831
Married, spouse present	0.472	0.645	0.402	0.596
Married, no spouse present	0.161	0.095	0.179	0.094
Divorced/widowed/sep	0.089	0.079	0.081	0.058
Never married	0.278	0.180	0.337	0.253
None	0.077	0.103	0.081	0.034
Elementary	0.110	0.074	0.134	0.056
Junior high	0.184	0.202	0.217	0.251
High school	0.290	0.245	0.337	0.310
College	0.244	0.292	0.200	0.297
Higher education	0.094	0.085	0.031	0.052
Europe	0.083	0.137	0.066	0.161
Asia	0.240	0.408	0.150	0.347
America	0.254	0.204	0.292	0.268
Africa	0.053	0.144	0.051	0.153
Mexico	0.370	0.102	0.441	0.068
Sample size	5,335	4,129	3,244	1,269

Australia and Canada are excluded from the analysis

immigrants are recorded in the ACS data. We further assume that the status of an immigrant is a choice variable determined by the following model:

$$Y = \begin{cases} 1 \text{ if } Y^* > 0 \\ 0 \text{ if } Y^* \leq 0 \end{cases}$$

where

$$Y^* = x\beta + \epsilon \tag{1}$$

with $\epsilon \sim N(0, \sigma_\epsilon^2)$. The set of covariates x includes variables that are assumed to be relevant in determining the relative gain of immigrating legally in the USA.

Since Y^* is the latent unobserved variable that determines the decision to be legal or illegal, we can write the probability of such choice as

$$P(Y = 1|x) = P(Y^* > 0|x) = P(\epsilon > -x\beta) = 1 - P(-x\beta) \tag{2}$$

and, assuming that $\epsilon \sim N(0, \sigma_\epsilon^2)$,

$$\begin{aligned} P(Y = 1|x) &= 1 - P(-x\beta) = 1 - F(-x\beta) \\ &= 1 - \int_{-\infty}^{\frac{-x\beta}{\sigma_\epsilon}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 1 - \Phi(-x\beta) = \Phi(x\beta) \end{aligned} \tag{3}$$

The total number of legal immigrants in the population is therefore given by

$$q = \int \Phi(x\beta) dF(x) \tag{4}$$

where $F(x)$ is the distribution function of the covariates x. We can also define the density function of the covariates x conditional on being observed in stratum 1 or 0. If they are observed in stratum 0, then their density function is the same as the unconditional density function, since stratum 0 randomly selects across all individuals:

$$p(x|s = 0) = f(x) \tag{5}$$

if instead, the observation comes from stratum 1, then the density function is given by

$$p(x|s = 1) = \frac{\Phi(x\beta) f(x)}{q} \tag{6}$$

Accordingly with Lancaster and Imbens (1996), we assume that the pooled sample is determined by a sequence of Bernoulli trials with parameter h unknown and independent on the other parameters of interest. Trials are repeated N times; each time if the trial results in a success, we randomly sample from the subpopulation with $Y = 1$; if instead results in a failure, we randomly select from the whole population. Therefore, in case of success, the sampled observation enters stratum $s = 1$ ("legal"), and in case of failure, the stratum $s = 0$. Then, h represents the probability that a randomly chosen observation from the pooled data belongs to stratum 1, while $1 - h$ to stratum 0.

Given these assumptions, we can write the joint density of stratum and covariates as follows:

$$g(x, s) = \left(\frac{h\Phi(x\beta) f(x)}{q}\right)^{s} \cdot \Big((1 - h) f(x)\Big)^{(1-s)} \tag{7}$$

The corresponding log likelihood function is therefore given by

$$\ell(\beta, h, q) = \sum_{n=1}^{N} \big[s_n \log[\Phi(x_n\beta) f(x_n)/q] + (1 - s_n) \log f(x_n)\big] + N_1 \log h + N_0 \log(1 - h) \tag{8}$$

In the form derived in Eq. 8 to be operative, the log likelihood needs that we know the density function $f(x)$. However, Lancaster and Imbens (1996) show that we can rewrite the likelihood greatly simplifying the problem and in particular eliminating the need to know the density of x. Define

$$R_1(x; \beta, q, h) = \frac{\frac{h}{q}\Phi(x\beta)}{\frac{h}{q}\Phi(x\beta) + 1 - h} \tag{9}$$

$$g(x) = \left[\frac{h}{q}\Phi(x\beta) + 1 - h\right] f(x) \tag{10}$$

$$R_0(x; \beta, q, h) = 1 - R_1(x; \beta, q, h), \tag{11}$$

then we can rewrite the log likelihood

$$\ell(\beta, h, q, \pi) = \sum_{n=1}^{N} \big[s_n \log[R_{1n}(x; \beta, q, h)] + (1 - s_n) \log R_{0n}(x; \beta, q, h)\big] + \sum_{n=1}^{N} \log g(x_n; \pi) \tag{12}$$

Lancaster and Imbens (1996) show that it is sufficient to maximize the first part of the likelihood function in order to obtain the maximum of the whole function. That is because the the four variables in the likelihood function are related by a functional relationship which is implicitly imposed by the maximization of the first part.[10] However, ignoring the second part of the likelihood makes it impossible to perform any inference because we do not know the actual value the likelihood function takes. Lancaster and Imbens (1996) solve this problem by showing that the first-order conditions for the maximization of the first part of the likelihood function can also be interpreted as a system of moments conditions leading to a generalized method of moments (GMM) estimation procedure. As such, it is possible to proceed with the GMM estimation and with inference using the covariance matrix from the GMM.

3.1 Accounting for different response rates

As the nonresponse rate is higher in the illegal immigrant population (20 % relative to 10 % in the legal immigrant population), we need to modify the model to reflect the undercounting of illegal immigrants relative to legals in the ACS. Because of undercounting, only a $\zeta_1 = 0.9$ portion of legal immigrants will enter the sample, while the portion of illegal will be only $\zeta_2 = 0.8$.[11] Given these proportions, unconditionally on other characteristics, the probability that a randomly chosen observation from the ACS belongs to a legal immigrant is given by

$$P(Y = 1|s = 0) = \frac{\zeta_1 P(Y = 1)}{\zeta_1 P(Y = 1) + \zeta_2 (1 - P(Y = 1))} = \frac{\zeta_1 q}{\zeta_1 q + \zeta_2 (1 - q)} \tag{13}$$

This implies that the density function of an observation from the ACS is given by

$$p(x|s = 0) = \frac{\zeta_1 q}{\zeta_1 q + \zeta_2 (1 - q)} \frac{\Phi(x\beta)}{q} f(x) + \frac{\zeta_2 (1 - q)}{\zeta_1 q + \zeta_2 (1 - q)} \frac{1 - \Phi(x\beta)}{1 - q} \tag{14}$$

$$f(x) = [\xi_1(q)\Phi(x\beta) + \xi_2)(1 - \Phi(x\beta))]f(x) \tag{15}$$

where $\xi_i(q) = \frac{\zeta_i}{\zeta_1 q + \zeta_2 (1-q)} = \frac{\zeta_i}{\zeta_2 + (\zeta_1 - \zeta_2) q}$, or defining $\zeta = \zeta_1 - \zeta_2$, $\xi_i(q) = \frac{\zeta_i}{\zeta_2 + \zeta q}$. Accordingly, (9)–(11) rewrite

$$R_1(x; \beta, q, h) = \frac{\frac{h}{q}\Phi(x\beta)}{\frac{h}{q}\Phi(x\beta) + (1 - h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1 - \Phi(x\beta))]} \tag{16}$$

[10] See Lancaster and Imbens (1996) page 149 for a discussion on this point.

[11] Tables 18 to 23 in Appendix 4 report sensitivity analysis with different assumptions on the undercounting rates for legal and illegal immigrants.

$$g(x) = \left[\frac{h}{q}\Phi(x\beta) + (1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1-\Phi(x\beta))]\right] f(x) \quad (17)$$

$$R_0(x;\beta,q,h) = 1 - R_1(x;\beta,q,h) \quad (18)$$

Taking the derivative of the log likelihood function with respect to β, we have, for the single observation,

$$\frac{\partial\ell(\beta,h,q)}{\partial\beta} = -R_1' \frac{s-R_1}{R_1(1-R_1)} \quad (19)$$

Defining N the numerator of R_1 and D the denominator, we have

$$R_1' = \frac{N'}{D} - R_1\frac{D'}{D} = \phi(x\beta)'\frac{1}{D}\left\{\frac{h}{q} - R_1[\frac{h}{q} + (1-h)(\xi_1 - \xi_2)]\right\} \quad (20)$$

where $\phi(x\beta) = \frac{\partial\Phi(x\beta)}{\partial\beta}$ is a $1\times\ k$ vector, k being the dimension of the vector β. Therefore,

$$R_1' = \phi(x\beta)'\frac{1}{D}\left\{(1-R_1)\left[\frac{h}{q} - R_1(1-h)(\xi_1-\xi_2)\right]\Big]\right\} \quad (21)$$

therefore,

$$\frac{\partial\ell(\beta,h,q)}{\partial\beta} = -\phi(x\beta)'(s-R_1)\left\{\frac{\frac{h}{q}}{DR_1} - \frac{(1-h)(\xi_1-\xi_2)}{D(1-R_1)}\right\} \quad (22)$$

or

$$\frac{\partial\ell(\beta,h,q)}{\partial\beta} = -\phi(x\beta)'(s-R_1)\left\{\frac{1}{\Phi(x\beta)} - \frac{(1-h)(\xi_1-\xi_2)}{(1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1-\Phi(x\beta))]}\right\} \quad (23)$$

Taking the derivative of the log likelihood function with respect to q, we have, for the single observation

$$\frac{\partial\ell(\beta,h,q)}{\partial q} = -\frac{s-R_1}{1-R_1}\left(\frac{1}{q} + \frac{D'}{D}\right) \quad (24)$$

where D is the denominator of R_1. Therefore, since

$$D' = -\frac{1}{q}\frac{h}{q}\Phi(x\beta) + (1-h)[\xi_1'(q)\Phi(x\beta) + \xi_2'(q)(1-\Phi(x\beta))]] \quad (25)$$

where

$$\xi_i' = -\frac{\zeta_i\zeta}{(\zeta_2+\zeta q)^2} = -\xi_i(q)\frac{\zeta}{\zeta_2+\zeta q} \quad (26)$$

therefore,

$$D' = -\frac{1}{q}\frac{h}{q}\Phi(x\beta) - (1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1-\Phi(x\beta))]]\frac{\zeta}{\zeta_2+\zeta q} \quad (27)$$

so that

$$\frac{D'}{D} = -\frac{1}{q}R_1 - (1 - R_1)\frac{\zeta}{\zeta_2 + \zeta q} \tag{28}$$

therefore,

$$\frac{\partial \ell(\beta, h, q)}{\partial q} = -\frac{s - R_1}{1 - R_1}\left(\frac{1}{q} + \frac{D'}{D}\right) = -\frac{s - R_1}{1 - R_1}\left(\frac{1}{q} - \frac{\zeta}{\zeta_2 + \zeta q}\right)(1 - R_1) \tag{29}$$

or

$$\frac{\partial \ell(\beta, h, q)}{\partial q} = -\left(\frac{1}{q} - \frac{\zeta}{\zeta_2 + \zeta q}\right)(s - R_1) \tag{30}$$

Equations (23) and (30) are the (single observation) equivalent of (3.6) and (3.7) in Lancaster and Imbens (1996), and (3.8) remains exactly the same in the modified model. The GMM interpretation of Lancaster and Imbens represented by the system of equations in (3.9) of their article needs to be modified accordingly, and is given by

$$\begin{aligned}
\psi_1(\beta, h, q, s, x) &= -\phi(x\beta)'(s - R_1)\left\{\frac{1}{\Phi(x\beta)} - \frac{(1-h)(\xi_1(q) - \xi_2(q))}{(1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1 - \Phi(x\beta))]}\right\}\\
\psi_2(\beta, h, q, s, x) &= -\left(\frac{1}{q} - \frac{\zeta}{\zeta_2 + \zeta q}\right)(s - R_1)\\
\psi_3(\beta, h, q, s, x) &= h - R_1
\end{aligned} \tag{31}$$

4 Results

We start by presenting results for the probability of being legal or illegal in the USA conditional on observed characteristics X. As a first formal estimation of the impact of personal characteristics on the propensity of being a legal or illegal immigrant, these results represent the paper's main contribution.

4.1 Estimating the probability of being legal/illegal

Table 3 shows the results from the modified probit model described in Section 3. The probability fitted by the model measures the likelihood that each observation belongs to a legal immigrant. We provide two sets of results, "unadjusted" and "adjusted"; in the former case, the sampling weights do not account for different nonresponse rates between legal and illegal aliens in the USA, while in the latter, they account for nonresponse rates of 2.5 % for legals and 10 % for illegals. We focus our discussion and subsequent applications on results that use the adjusted weights, but the two sets of results are very similar, and moreover, in Table 11 in Appendix 1, we repeat the analysis using unadjusted sampling weights.

The numbers represent probit coefficients; marginal effects would be scaled down by a positive factor of $\varphi(X\beta)$. All independent variables are categorical dummies,

Table 3 Probit results: conditional probability of being a legal immigrant from NIS and ACS 2007

	Unadjusted		Adjusted	
	Estimate	Std. error	Estimate	Std. error
Constant	−2.4244	0.1773	−2.4567	0.1732
Female	0.1215	0.0489	0.1198	0.0479
Age	0.0515	0.0041	0.0508	0.0041
Elementary	0.6491	0.1597	0.6407	0.1564
Junior high	1.0955	0.1731	1.0822	0.1692
High school	0.5520	0.1351	0.5448	0.1323
College degree	0.3450	0.1319	0.3406	0.1290
Higher education	−0.0989	0.1444	−0.0978	0.1413
Married spouse present	0.1477	0.0554	0.1462	0.0543
Married spouse not present	−0.5245	0.0866	−0.5176	0.0852
Mex*<elementary	−0.7546	0.1948	−0.7442	0.1904
Mex*elementary	−1.7213	0.1815	−1.6985	0.1785
Mex*junior high	−2.1610	0.2083	−2.1335	0.2050
Mex*high school	−1.6654	0.1542	−1.6434	0.1520
Mex*college degree	−1.6102	0.2121	−1.5892	0.2105
Mex*higher education	−1.0224	0.3891	−1.0085	0.3865
America	−0.8648	0.1095	−0.8529	0.1068
Africa	0.5357	0.1450	0.5300	0.1396
Asia	−0.3633	0.0942	−0.3577	0.0913
q	0.4175	0.0476	0.3945	0.0473
LogLik	−3,574.21		−3,574.22	
No. of obs	9,464		9,464	

Data: NIS 2003 and ACS 2007, excluding students, Canadians and Australians, and including professionals. "Unadjusted" uses normalized sampling weights provided by each survey. "Adjusted" rescales sampling weights to account for differences in nonresponse rates. Reference category: below elementary education (0 to 4 years of school), single, European. Education categories: (1) below elementary = none to grade 4 (base); (2) elementary = grades 5 to 8; (3) junior high = grades 9 to 12, no diploma; (4) high school diploma; (5) college = postsecondary education up to Bachelor's; (6) higher education = Master's, professional degrees, Ph.D.

except for age which is measured in years, and all coefficients with the exception of education beyond Bachelor's are statistically significant.

The positive coefficient on the gender variable indicates that women are less likely to be illegally present in the USA than men; women may be more risk averse than men, may have dependent children who make illegal immigration more costly, or the return from illegal immigration could be lower for women than for men. Relative to being single, being married with the spouse present increases the probability to be legally in the USA, while being married and living without the spouse decreases it. Age and education have the expected effect on the probability to be legal: older immigrants and more educated immigrants are more likely to be legal. The only exception is for the highest category of education: all else equal, immigrants with Master's, professional degrees, or Ph.D.'s seem to be less likely to be legal—although the coefficient is not statistically significant. The negative impact may be an indication that, despite our best efforts, we may still misclassify some visa holders as illegal immigrants—especially the professional H-type visas, the largest visa category not a priori excluded from the analysis. With this in mind, we report in Section 5.1 sensitivity analysis from dropping professional occupations from the analysis.

Due to its proximity, Mexico is the biggest source of illegal immigration into the USA, and the determinants of legal/illegal immigration may differ for Mexicans compared to other source countries. To this extent, we estimate a separate set of interaction dummies between schooling levels and Mexican origin. Indeed, we find that (i) at all levels of education, Mexicans are more likely to be illegal in the USA than immigrants from other countries, and (ii) in terms of education, for Mexicans, the opposite holds true: the more educated an immigrant, the more likely she/he is to be undocumented in the USA.[12]

Relative to baseline Europe, the continent of origin coefficients are significant and negative, except for Africa, indicating that African-born immigrants are more likely to be legally in the USA than Europeans, while Asians and Latin Americans are more likely to be illegally there.[13]

Lastly, q, the unconditional probability of being legal, seems to be slightly underestimated at 0.395 or 0.418, depending on the specification (correcting or not for different nonresponse rates). We can get slightly higher estimates of q, more in line with DHS projections, from the sensitivity analysis where we drop professionals and thus reduce some of the noise coming from visa observations.

4.2 Statistics on legal and illegal immigrants inferred from ACS

From the probit model discussed above, we can compute for each observation the probability that an individual has legal or illegal status given her/his characteristics. We use these probabilities as weights to infer the distribution of certain characteristics in the legal (illegal) subpopulation of immigrants. We start by investigating discrete categorical variables Z, and we focus on the first moment, the mean.[14]

As a measure of how well our methodology performs, we compare the means predicted from our model for variables in the illegal population with benchmark statistics

[12] For Mexicans, the total effect of schooling on the propensity of legal immigration is the sum of the effect of education (relative to elementary) plus the effect of Mex∗education (relative to elementary). In our specification, it should be computed as the difference between the "Mex∗education" interaction coefficient and the coefficient of "Mex∗ <elementary (which is -0.74 in the adjusted and $0 - .75$ in the unadjusted specifications). This overall effect is negative and small.

[13] Canadian observations are left out of the estimation, so America includes all of Central and South America except for Mexico, which is estimated alone.

[14] The mean in the overall immigrant population is a weight between the legal and illegal means, with weights given by the unconditional probability of being legal (q) and illegal ($1 - q$).

reported by the DHS Office of Immigration Statistics for 2007, as in Hoefer et al. (2008).[15]

We focus on immigrants who migrated on and after year 2000; since our procedure relies on data on relatively recent immigrants, we feel more confident in extrapolating the results to a sample that is closer to the one used for our estimates.[16] Means for education categories, marital status, country of origin, and US state of current residence are reported in Table 4. The first column, (i) "All ACS", reports the means in the entire immigrant population using the ACS sample from 2007. The next two columns focus on the illegal subpopulation: (ii) "Illegal ACS" uses our methodology of computing means in the illegal subpopulation by applying the weights $1 - \omega_i(X_i)$, while (iii) "Illegal DHS" has benchmark estimates coming from the DHS benchmark statistics on undocumented immigrants. The last two columns provide the means in the legal immigrant subpopulation: (iv) "Legal ACS" computed by applying our legal probability weights ω_i to the ACS data, and (v) "NIS" which reports summary statistics from the NIS data on legal green card holders, and can thus be seen as our own benchmark for the legal immigrant subpopulation.

Note that we do not need that the observed characteristics X involved in the determination of the legal/illegal probabilities as reported in Table 3 match one-on-one with the variables whose mean we compute here. For instance, we did not use the state of residence variable when computing the legal/illegal weights because the state information was not reliable in the NIS data, where we know the state where the green card was mailed, which is not necessarily the state where the immigrant resides in 2007 (or even 2003 for that matter).

In terms of model fit, our statistics are a very reasonable match for the benchmark. For instance, there are 36.5 % Mexican immigrants in the general population of immigrants (ACS); our procedure makes this fraction go up to 44 % in the illegal subpopulation, moving it closer to the 59 % illegal Mexican immigrants reported by the DHS (Hoefer et al. 2008). Our procedure estimates 44 % women in the illegal subpopulation, same as in the DHS estimates, and 56.6 % women in the legal immigrant subpopulation, compared to 51.4 % in the benchmark NIS. (In the general immigrant population in ACS, the percentage of women is 46.7 %). For Asians, the illegal percentage decreases from 25.3 % in the overall immigrant population to

[15] One issue about comparing our estimates with the ones from DHS is that we drop several observations as we implicitly assume that although they belong to non LPR, they belong, with very high probability, to legally resident immigrants. In particular, students comprise the bigger share of dropped immigrants. The DHS estimates include students as well. In Appendix 2, we report three tables on students. Table 12 reports the results assuming that students are all illegal, while the weights for the rest of the immigrants are extrapolated using our estimation results in Table 3. The following table assumes that students are all legal (Table 13). In the last, Table 14, students are given the same weights as in the rest of all immigrants. From Tables 12, 13, and 14, it is clear that the assumption that generates statistics closer to the DHS is the one we implicitly make by dropping students from our analysis, that is, that students are most likely legally present in the USA.

[16] Note that the DHS estimates are on the entire population of immigrants after 1980. In Tables 15, 16, and 17 in Appendix 3, we report estimates made using data on all immigrants from 1980; while the results are slightly different, they do not change qualitatively. Indeed, our estimates using different samples are in line with the estimates reported by DHS on year 2000 and year 2007.

Table 4 Legal and illegal distributions from ACS 2007

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0717	0.0718		0.0716	0.0845
Elementary	0.1558	0.1691		0.1051	0.1277
Junior	0.1341	0.1362		0.1261	0.1517
High school	0.2921	0.2929		0.2891	0.2705
Some college	0.2484	0.2291		0.3217	0.2634
Higher education	0.0979	0.1009		0.0865	0.1023
Married with sp	0.4307	0.3975		0.5571	0.8063
Married no sp	0.1489	0.1584		0.1127	0.0475
Single	0.4196	0.4431		0.3300	0.1462
European	0.0926	0.0678	0.0200	0.1870	0.1427
Asian	0.2528	0.2163	0.1200	0.3915	0.3282
American	0.2386	0.2470	0.2400	0.2065	0.2548
African	0.0513	0.0282	0.0200	0.1394	0.1000
Mexican	0.3647	0.4406	0.5900	0.0757	0.1743
Sex	0.4667	0.4407	0.4400	0.5657	0.5140
California	0.2128	0.2145	0.2400	0.2064	0.0000
Texas	0.1038	0.1115	0.1400	0.0746	0.0000
Florida	0.0974	0.0953	0.0800	0.1052	0.0000
Arizona	0.0328	0.0364	0.0500	0.0192	0.0000
New York	0.0944	0.0889	0.0500	0.1156	0.0000

Data: ACS 2007, excluding students, Canadians, Australians, and immigrants who migrated prior to year 2000, and including professionals. Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer et al. (2008). NIS = statistics on legal green card holders from the 2003 NIS (also used in computing the weights). Estimation using the legal/illegal probability weights from ACS 2007, Table 3 (sampling weights rescaled to account for differences in nonresponse rates). For a description of education categories, see footnotes to Table 3

21.6 % in the illegal population, whereas the DHS benchmark is 12 %; the legal frequency increases to 39 %, which is closer to the NIS 32.8 % benchmark.

In the probit analysis, a more educated immigrant was more likely to be legal. We see some of the same effect here: postsecondary education is more prevalent (higher mean) in the legal subpopulation, while professional degrees, whose effect was statistically insignificant in the probit, have the same means in the legal and illegal subpopulations. Despite having positive coefficients in the probability model, we do not see lower levels of education (such as junior high or high school) more frequently in the legal subpopulation. We believe that this is due to Mexican immigrants for whom the probability to be legal is negatively related to education; their presence brings down the mean of education at lower levels of education within the legal subpopulation. Like in our previous discussion of probit coefficients, married individuals with spouse present are more frequent in the legal population, while single individuals or married with spouse absent are more frequent in the illegal population. As mentioned before, we did not use state of residence as a predictor for legal/illegal probabilities. In terms of forecast means, California, Texas, and Arizona have more illegal immigrants than legal, New York has more legal immigrants, while Florida has about the same.

5 Sensitivity analysis

5.1 Sensitivity to dropping professionals from ACS 2007

This section repeats the previous analysis with one difference: professionals are dropped from the estimation, to further minimize the chance of ACS including

Table 5 Probit results: conditional probability of being a legal immigrant from NIS and ACS 2007; excluding professionals

	Uncorrected		Corrected	
	Estimate	Std. error	Estimate	Std. error
Constant	−2.0166	0.3388	−2.0532	0.3303
Female	−0.0071	0.0941	−0.0070	0.0922
Age	0.0489	0.0071	0.0483	0.0070
Elementary	0.8234	0.2762	0.8134	0.2723
Junior high	1.3615	0.3076	1.3471	0.3026
High school	0.7078	0.2386	0.6994	0.2355
College and higher	0.5987	0.2374	0.5914	0.2340
Married spouse present	0.1772	0.1030	0.1755	0.1011
Married spouse not present	−0.6536	0.1572	−0.6462	0.1552
Mex*<elementary	−1.4462	0.4188	−1.4285	0.4118
Mex*elementary	−2.2500	0.3646	−2.2227	0.3571
Mex*junior high	−2.5591	0.4047	−2.5301	0.3965
Mex*high school	−1.9233	0.3048	−1.9001	0.2977
Mex*college and higher	−2.1489	0.4086	−2.1234	0.4041
America	−1.1044	0.2469	−1.0912	0.2385
Africa	0.4732	0.3023	0.4692	0.2882
Asia	−0.4731	0.2264	−0.4664	0.2169
q	0.5206	0.0866	0.4969	0.0866
LogLik	−1,395.51		−1,395.51	
No. of obs	4,513		4,513	

Data: NIS 2003 and ACS 2007, excluding students, Canadians and Australians, and excluding professionals. “Unadjusted” uses normalized sampling weights provided by each survey. “Adjusted” rescales sampling weights to account for differences in nonresponse rates. Reference category: below elementary education (0 to 4 years of school), single, European. For a description of education categories, see footnotes to Table 3 immigrants with legal temporary status. Table 5 reports the results of the probit estimation and Table 6 the corresponding means in the legal and illegal subpopulations, once professionals are dropped from the analysis. Here, we report results from analysis which re-weights for nonresponse rates, while Appendix 1 contains sensitivity results when nonresponse rates are not accounted for.

The probit results are very much in line with those from the analysis on the entire sample which were reported in Table 3. One notable exception is the higher fraction of legal immigrants in the population, q, which is now 0.521 and respectively 0.497, depending on whether weights are adjusted or not for differential nonresponse rates in the surveys. This higher q indicates that at least some of the professionals must have been identified as illegal previously. The two highest education categories are now grouped, because after dropping professional occupations, their size has reduced considerably. At the same time, we see a much higher likelihood for an educated

Table 6 Legal and illegal distributions from ACS 2007, *excluding professionals*. Weights from 2007 ACS

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0717	0.0814		0.0465	0.0845
Elementary	0.1558	0.1824		0.0863	0.1277
Junior	0.1341	0.1408		0.1167	0.1517
High school	0.2921	0.3012		0.2683	0.2705
Some college	0.2484	0.2155		0.3343	0.2634
Higher education	0.0979	0.0788		0.1479	0.1023
Married with sp	0.4307	0.3743		0.5780	0.8063
Married no sp	0.1489	0.1683		0.0984	0.0475
Single	0.4196	0.4564		0.0984	0.1462
European	0.0926	0.0495	0.0200	0.2052	0.1427
Asian	0.2528	0.1908	0.1200	0.4148	0.3282
American	0.2386	0.2542	0.2400	0.1978	0.2548
African	0.0513	0.0209	0.0200	0.1310	0.1000
Mexican	0.3647	0.4846	0.5900	0.0511	0.1743
Sex	0.4667	0.4413	0.4400	0.5330	0.5140
California	0.2128	0.2178	0.2400	0.1999	0.0000
Texas	0.1038	0.1162	0.1400	0.0716	0.0000
Florida	0.0974	0.0960	0.0800	0.1011	0.0000
Arizona	0.0328	0.0387	0.0500	0.0176	0.0000
New York	0.0944	0.0860	0.0500	0.1165	0.0000

Data: ACS 2005 to 2007, excluding students, Canadians and Australians, excluding professionals. Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer et al. (2008). NIS = statistics on legal green card holders from the 2003 NIS. Estimation using the legal/illegal probability weights from ACS 2007 (no professionals); sampling weights rescaled to account for differences in nonresponse rates. For a description of education categories, see footnotes to Table 3 individual of being legal. In fact, compared to the base analysis, all education categories indicate a slightly higher probability of being legal relative to individuals with no education. Also, the coefficient on women has become negative, indicating that, once professionals are excluded, women are less likely to be legal than men.

From the distribution of covariates reported in Table 6, we can see that excluding professionals results in a smaller gap between the number of illegals forecast by our procedure and the benchmark DHS number, as well as for statistics for country of origin and US state of residence. In particular, we predict a fraction of illegal Mexicans closer to the one reported by the DHS. We also get a better forecast fit for gender and some education categories such as high school.

6 Immigrants' human capital and other characteristics

6.1 Returns to schooling and experience

Our methodology allows us to determine the conditional probability of each immigrant in the cross section to be a legal or illegal resident. We use these probability weights in a Mincer wage regression to investigate the comparative returns to schooling and experience for legal and illegal immigrants. We consider in the wage regression all the immigrants from the 2005 to 2007 waves of the ACS who have immigrated since 2001 (Table 7). The conditional probability weights of being a legal or illegal immigrant are computed using ACS 2007 for immigrants who reported 2003 as their entry time in the USA. The dependent variable is log wages, and thus the OLS coefficients can be approximated as percentage effects, except at larger values of the coefficients where the exact percentage values need to be computed as $\exp(\beta) - 1$; we refer here to log points. While the findings are not surprising, this is a very relevant exercise because it quantifies the magnitude of human capital returns within the legal and illegal populations. We do the analysis separately by gender.

All else equal, being an illegal male immigrant brings a substantive wage penalty of 57 log points relative to a legal immigrant; for females, the penalty is 44 log points. Potential experience, which we construct as age-schooling-6, has the expected small positive effect on wages, at a decreasing rate. Wages grow between 4 to 6 % each year, as indicated by survey year dummies.

Relative to elementary education, having some high school but no diploma ("junior high") seems to hurt legal immigrants, for whom the return is negative: a penalty of 18.6 log points for men and 12.5 log points for women. This is no longer true for illegal immigrants, especially for illegal male immigrants who get a positive return of about 5 % from having more than elementary education. High school diplomas seem to have no significant impact except for illegal male immigrants who get positive returns from having graduated high school.

Having a college degree has large significant returns for both immigrant men and women relative to uneducated immigrants. The return is similar for legal and illegal women immigrants, but there is a penalty for illegal immigrant men. For them, the return to college, while still positive, is much smaller than for their legal immigrant counterparts: 43 log points compared to 66 log points. The same story holds for

Table 7 Returns to legal status and education from ACS 2007

	Males		Females	
	Estimate	Std. error	Estimate	Std. error
Constant	2.3886	0.0569	2.1718	0.0592
Junior high	−0.1860	0.0623	−0.1250	0.0646
High school	−0.0025	0.0586	0.0745	0.0596
College	0.6657	0.0596	0.4064	0.0613
Higher education	1.0165	0.0758	0.8207	0.0875
Illegal	−0.5729	0.0598	−0.4401	0.0649
Junior high/illegal	0.2397	0.0675	0.1552	0.0768
High school/illegal	0.1466	0.0635	0.0335	0.0705
College/illegal	−0.2333	0.0663	0.0694	0.0739
Higher education/illegal	−0.0558	0.0877	0.0864	0.1073
Experience	0.0242	0.0014	0.0185	0.0020
Experience^2	−0.0005	0.0000	−0.0005	0.0000
Survey year 06	0.0352	0.0079	0.0360	0.0112
Survey year 07	0.0592	0.0077	0.0750	0.0109
R^2		0.2440		0.1800
No. of obs		43,443		25,057

Data: ACS 2007, individuals who have immigrated since 2001, excluding students, Canadians and Australians, and including professionals. Estimation using the legal/illegal probability weights from ACS 2007, Table 3; sampling weights rescaled to account for differences in nonresponse rates. "Illegal" defined as the conditional probability for each observation to be legal. For a description of education categories, see footnotes to Table 3

postgraduate education, where returns are very large for legal immigrants, almost double than for college. The penalty for illegal immigrants in this case is smaller and not significant, for both men and women.

6.2 Sensitivity to using a different set of legal/illegal immigrant weights, excluding professionals

For a sensitivity check, we repeat the wage regression analysis using a different set of legal/illegal conditional probabilities, obtained from ACS 2007 excluding professionals. The results are reported in Table 8 and tell a similar story: there is an overall wage penalty from being an illegal immigrant relative to a legal one, although the penalty is somewhat reduced as the average educational attainment is overall lower; on top of it, the penalty is even higher for college-educated men, but not for other education and demographic categories. The returns to college, and especially to postgraduate education are very large and remain positive even for illegal immigrants. Relative to little or no education, all other levels of education receive a premium, except for immigrants with junior high education, who fare worse on average. The illegal

Table 8 Returns to legal status and education from ACS 2007 using illegal immigrant weights from ACS 2007 excluding professionals

	Males		Females	
	Estimate	Std. error	Estimate	Std. error
Constant	2.2001	0.0450	2.0576	0.0512
Junior high	−0.0608	0.0492	−0.0554	0.0558
High school	0.1141	0.0460	0.1325	0.0513
College	0.7250	0.0459	0.4860	0.0516
Higher education	1.0195	0.0568	0.8106	0.0711
Illegal	−0.3668	0.0463	−0.3113	0.0546
Junior high/illegal	0.1033	0.0542	0.0717	0.0668
High school/illega	0.0141	0.0506	−0.0410	0.0610
College/illegal	−0.3724	0.0525	−0.0620	0.0631
Higher education/illegal	−0.0747	0.0706	0.1024	0.0914
Experience	0.0221	0.0014	0.0174	0.0020
Experience^2	−0.0004	0.0000	−0.0004	0.0000
Survey year 06	0.0357	0.0079	0.0364	0.0112
Survey year 07	0.0602	0.0077	0.0755	0.0109
R^2		0.2475		0.1811
No. of obs		43,443		25,057

Data: ACS 2007, individuals who have immigrated since 2001, excluding students, Canadians and Australians, and excluding professionals. Estimation using the legal/illegal probability weights from ACS 2007, Table 3; sampling weights rescaled to account for differences in nonresponse rates. "Illegal" defined as the conditional probability for each observation to be illegal, $1 - \omega_i(X_i)$. For a description of education categories, see footnotes to Table 3

immigrant penalty varies by education categories, with a big difference between men and women. For men, being illegal with junior high education brings return, while being illegal college-educated carries a penalty. For women, the interaction terms are not statistically significant.

These results which illustrate that, as expected, illegal immigrants suffer a wage penalty due to their status, can be further seen from nonparametric wage distribution plots: the wage distribution for legal immigrants presents a higher mean and more skewness to the right compared to the wage distribution for illegal immigrants (Fig. 1).

While there certainly appears to be a large penalty for illegal status, heterogeneous depending on eduction and gender, higher educated immigrants still get a substantive overall premium to their education. For instance, illegal immigrants with a postgraduate education will have a benefit from education of about 40 to 50 percentage points higher than uneducated legal immigrants. From a policy standpoint, this may warrant further thought into the welfare implications of a skill-selective immigration policy like the ones employed by Australia or Canada.

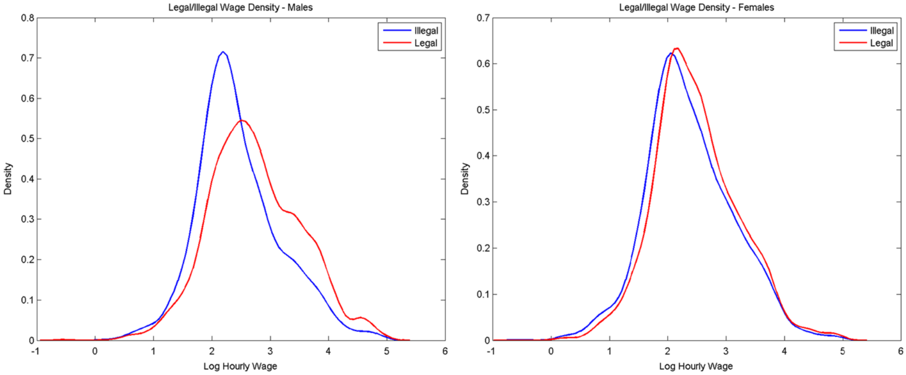

Fig. 1 Log-wage densities—legal and illegal immigrants

6.3 Sector of activity

Table 9 shows the distribution of immigrants across sector of activity. The first column indicates the results for all immigrants, while the second and third column show the results for legal and illegal immigrants. The distributions of legal and illegal immigrants do not change substantially compared to the distribution for all immigrants, except for some notable sectors. In agriculture, construction, and, to a slightly lesser extent, recreational and entertainment services, it seems that the concentration of illegal immigrants is much more significant than in other sectors. On the contrary, in sectors like health care and retail trade, there seems to be a prevalence of legal immigrants. To some extent, those are the results that we would expect, as there is plenty of anecdotal evidence that illegal immigrants concentrate in low-skill jobs, especially in constructions and agriculture (therefore providing further reassurance for the good performance of our estimated probabilities). However, with our estimates, we can go further and look at less obvious results such as unemployment. In this case, we notice that the unemployed are more likely to be legal immigrants rather than illegal. This also can have an easy explanation in economic terms as illegal immigrants may be willing to accept lower paid jobs as it may be hard for them to qualify for unemployment benefits and easier to disappear in the underground economy.

6.4 Fertility

Another set of statistics we can compute is the fertility rate for women aged between 15 and 49, by education and by legal status. Table 10 reports these estimates. The overall fertility rates are higher compared to standard statistics because of the sample selection. In particular, the fact that we restrict the analysis to recent immigrants

Table 9 Sector of activity from ACS 2007 using illegal immigrant weights from ACS 2007 excluding professionals

	All	Legal	Illegal
Agriculture	0.0314	0.0110	0.0368
Extraction	0.0015	0.0017	0.0015
Utilities	0.0014	0.0018	0.0013
Construction	0.1460	0.0780	0.1642
Manufacturing	0.1036	0.1046	0.1033
Wholesale trade	0.0262	0.0262	0.0262
Retail trade	0.0711	0.0913	0.0657
Transportation	0.0208	0.0299	0.0183
Information	0.0111	0.0143	0.0103
Finance	0.0265	0.0349	0.0243
Professional services	0.1094	0.1012	0.1116
Education	0.0277	0.0383	0.0249
Health care	0.0463	0.0768	0.0382
Individual and family services	0.0039	0.0067	0.0031
Recreation and entertainment	0.1226	0.1099	0.1260
Other services	0.0529	0.0606	0.0509
Public administration	0.0041	0.0064	0.0035
Military	0.0010	0.0012	0.0009
Unemployment	0.0127	0.0132	0.0126

Data: ACS 2005–2007, individuals who have immigrated since 2001, excluding students, Canadians and Australians, and excluding professionals. Estimation using the legal/illegal probability weights from ACS 2007, Table 3; sampling weights rescaled to account for differences in nonresponse rates. “Illegal” defined as the conditional probability for each observation to be illegal, $1 - \omega_i(X_i)$. For a description of education categories, see footnotes to Table 3

increases the fertility for all cohorts of women and, as such, the total fertility rate as reported in the table.[17]

Among illegal women, fertility is higher than among legal ones. In particular, among legal immigrants, we see the typical decreasing relationship between

[17] The total fertility rate is calculated as the average of the fertility rates for each cohort, times 5 (we consider seven 5-year cohorts). The fertility rate for each cohort is calculated as the percentage of women who have had one live birth in the last 12 months prior the interview. We restrict to immigrants who immigrated after the year 2000. Some women who intended to migrate and have children are more likely to have waited until the successful migration to have children. We also restrict the sample to nonstudents, which increases the fertility rate of the younger women left in the sample.

Table 10 Fertility rates from ACS 2007 using illegal immigrant weights from ACS 2007 excluding professionals

Education	All	Legal	Illegal
Elementary	4.59	4.48	4.57
Middle school	4.67	3.85	4.72
High school and higher	3.75	3.2	3.76

Data: ACS 2005–2007, individuals who have immigrated since 2001, excluding students, Canadians and Australians, and excluding professionals. Estimation using the legal/illegal probability weights from ACS 2007, Table 3; sampling weights rescaled to account for differences in nonresponse rates. "Illegal" defined as the conditional probability for each observation to be illegal, $1 - \omega_i(X_i)$. For a description of education categories, see footnotes to Table 3

education and fertility, which we do not see monotonically among all immigrants. Among illegal instead, fertility seems to be higher at middle levels of education than at the lowest or the highest level, and it remains high at high levels of education. We do not have an immediate or straightforward explanation for the differences, but we do believe that our results pose more questions to be addressed by future research.

7 Conclusion

In this paper, we have provided methodology to separate the legal and illegal immigrants from two random surveys in the USA. Using information on all US immigrants from ACS and information on legal US immigrants from NIS, we were able to identify a set of probability weights which, conditional on observed characteristics, can determine the likelihood for each individual to be a legal or an illegal immigrant, based on the observed characteristics. From a substantive point of view, we wanted to use this methodology in investigating what are the characteristics of legal versus illegal immigrants and whether the legal status of an immigrant has an impact on their human capital, wages, and returns to human capital.

Compared to legal immigrants, we have quantified to what extent illegal immigrants are more likely to be less educated, males, and married with spouse not present. These results are heterogeneous across education categories, country of origin (Mexico), and whether professional occupations have been included in the analysis. While illegal immigrants experience a large wage penalty compared to legal immigrants, returns to higher education remain large and positive compared to legal uneducated immigrants. The illegal immigrant penalty is heterogeneous across education categories and gender, with women experiencing a lower penalty for being illegal compared to men. We also find that the total fertility rate among illegal immigrant women is significantly higher than among legal ones, and that is particularly true for middle and higher educated women. Finally, looking at the sector of activity, we find that constructions is the sector that most attracts illegal immigrants and that most of the immigrants in agriculture are also illegal.

Future research can use the weights computed here in a variety of applications where it is interesting to differentiate between legal and illegal immigrants. Using the estimated probit coefficients β reported here and observed immigrant characteristics either from the ACS or from other microdata with information on immigrants, researchers can compute the probability weight for each observation to be a legal or illegal immigrant. These weights can be used in returns to human capital in wage estimations, like we did here, or in other analysis involving the immigrant population. Examples include studies which assess the net contribution of immigrants to welfare, such as Social Security, Unemployment Insurance, and other government transfer programs to which illegal immigrants contribute. Our weighting methodology can also apply to studies related to immigrants' access to education, income inequality, and intergenerational mobility and, more broadly, productivity in general. By using the conditional immigrant weights, the analysis can identify outcomes for legal and illegal immigrants not only at the mean, but also along the distribution of these outcomes, like in the wage densities example we presented here.

Furthermore, we can also use our weights to decompose the wage differential between legal and illegal immigrants at all quantiles of the earnings distribution. Following the density re-weighting methodology from Fortin et al. (2011), by applying the illegal probability weights to the NIS data, we can recover in a semi-parametric approach the counterfactual wage density for legal immigrants, had they had the characteristics of illegal immigrants, and decompose the legal/illegal immigrant wage gap.

Some caveats apply. We had to be extremely careful in how we treated immigrant visa holders, whom we could not directly observe, and were concerned not to misidentify as illegals. We believe that all the sensitivity analyses indicate that our approach was successful in that regard. Another caveat is that we focus on the 2003 flow data; as such, our methodology can generalize to other immigrant cohorts only to the extent that there have not been major demographic changes in the composition of legal versus illegal immigrant flows. If NIS releases subsequent waves of the survey, we can update the weights to reflect the experience of more recent immigrants.

We see as our main contribution the fact that we were able to use representative microdata to back out legal immigrant status out of personal characteristics and then predict the relative labor market performance of the two categories. Our methodology should be of interest to all researchers who need to make some inferences based on legal or illegal immigrant status.

Appendix 1: Using 2007 ACS with sampling weights not adjusted for differences in nonresponse rates

Table 11 Legal and illegal distributions from ACS 2007, including professionals, survey sampling weights

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0717	0.0718		0.0714	0.0845
Elementary	0.1558	0.1701		0.1047	0.1277
Junior	0.1341	0.1366		0.1250	0.1517
High school	0.2921	0.2930		0.2888	0.2705
Some college	0.2484	0.2276		0.3226	0.2634
Higher education	0.0979	0.1009		0.0875	0.1023
Married with sp	0.4307	0.3953		0.5570	0.8063
Married no sp	0.1489	0.1590		0.1129	0.0475
Single	0.4196	0.4447		0.3299	0.1462
European	0.0926	0.0665	0.0200	0.1860	0.1427
Asian	0.2528	0.2141	0.1200	0.3910	0.3282
American	0.2386	0.2471	0.2400	0.2082	0.2548
African	0.0513	0.0272	0.0200	0.1377	0.1000
Mexican	0.3647	0.4451	0.5900	0.0772	0.1743
Sex	0.4667	0.4392	0.4400	0.5648	0.5140
California	0.2128	0.2147	0.2400	0.2060	0.0000
Texas	0.1038	0.1120	0.1400	0.0748	0.0000
Florida	0.0974	0.0952	0.0800	0.1053	0.0000
Arizona	0.0328	0.0366	0.0500	0.0192	0.0000
New York	0.0944	0.0885	0.0500	0.1156	0.0000

Data: ACS 2005 to 2007, excluding students, Canadians and Australians, but including professionals. Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer et al. (2008). NIS = statistics on legal green card holders from the 2003 NIS. Estimation using the legal/illegal probability weights from ACS 2007, Table 3. No correction for differences in nonresponse rates. For a description of education categories, see footnotes to Table 3

Appendix 2: Students

This appendix presents the distribution of legal and illegal immigrants under different hypotheses about students. The first table shows the distribution of illegal and legal immigrants when we assume that students are all illegal, i.e., the distributions are calculated assigning a probability to be legal to students equal to 0. The second table assumes a probability to be legal equal to 1. This is the benchmark case, as what we implicitly assume taking students of the sample when we calculate the probabilities of being illegal. The difference is that in this case, the distribution of legal immigrants is changed accordingly to the fact that students are counted in this subpopulation. Notice that we do not include students in the probit step of the computation of the probabilities to be illegal as we believe that this would bias our estimation. Our estimation suggests that younger and less educated immigrants are more likely to be illegal; this, however, is probably not true for students that are younger because of their status as students and less educated because they are still acquiring education. The last table shows the distribution assuming that students are like all other immigrants. Clearly, the second table is the one that most approaches the DHS estimates. Therefore, our hypothesis that students are mostly legally present in the USA is the one that best helps to reproduce the aggregate DHS statistics.

Table 12 Legal and illegal distributions: with students assumed to be illegal (Pr(illegal) = 1)

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0473	0.0492		0.0339	0.0845
Elementary	0.1462	0.1556		0.0772	0.1277
Junior	0.1613	0.1655		0.1308	0.1517
High school	0.2719	0.2678		0.3019	0.2705
Some college	0.2728	0.2611		0.3578	0.2634
Higher education	0.1005	0.1008		0.0984	0.1023
Married with sp	0.3749	0.3395		0.6335	0.8063
Married no sp	0.1179	0.1242		0.0720	0.0475
Single	0.4907	0.5177		0.2942	0.1462
European	0.0956	0.0817	0.0200	0.1969	0.1427
Asian	0.2717	0.2579	0.1200	0.3724	0.3282
American	0.2311	0.2368	0.2400	0.1896	0.2548
African	0.0637	0.0481	0.0200	0.1773	0.1000
Mexican	0.3379	0.3755	0.5900	0.0637	0.1743
Sex	0.4626	0.4500	0.4400	0.5541	0.5140
California	0.2046	0.2084	0.2400	0.1773	0.0000
Texas	0.1021	0.1056	0.1400	0.0770	0.0000
Florida	0.0904	0.0902	0.0800	0.0915	0.0000
Arizona	0.0308	0.0329	0.0500	0.0159	0.0000
New York	0.0966	0.0935	0.0500	0.1190	0.0000

Table 13 Legal and illegal distributions: with students assumed to be legal (Pr(illegal) = 0)—benchmark

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0473	0.0651		0.0122	0.0845
Elementary	0.1462	0.1691		0.1008	0.1277
Junior	0.1613	0.1398		0.2040	0.1517
High school	0.2719	0.2962		0.2240	0.2705
Some college	0.2728	0.2285		0.3604	0.2634
Higher education	0.1005	0.1015		0.0986	0.1023
Married with sp	0.3749	0.3998		0.3256	0.8063
Married no sp	0.1179	0.1498		0.0548	0.0475
Single	0.4907	0.4494		0.5726	0.1462
European	0.0956	0.0679	0.0200	0.1505	0.1427
Asian	0.2717	0.2117	0.1200	0.3905	0.3282
American	0.2311	0.2437	0.2400	0.2062	0.2548
African	0.0637	0.0289	0.0200	0.1326	0.1000
Mexican	0.3379	0.4478	0.5900	0.1201	0.1743
Sex	0.4626	0.4366	0.4400	0.5142	0.5140
California	0.2046	0.2129	0.2400	0.1882	0.0000
Texas	0.1021	0.1119	0.1400	0.0827	0.0000
Florida	0.0904	0.0932	0.0800	0.0847	0.0000
Arizona	0.0308	0.0361	0.0500	0.0204	0.0000
New York	0.0966	0.0882	0.0500	0.1132	0.0000

Table 14 Legal and illegal distributions: with students with general weights (i.e., assumed to be similar to others)

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0473	0.0508		0.0274	0.0845
Elementary	0.1462	0.1593		0.0712	0.1277
Junior	0.1613	0.1627		0.1537	0.1517
High school	0.2719	0.2699		0.2838	0.2705
Some college	0.2728	0.2558		0.3695	0.2634
Higher education	0.1005	0.1016		0.0944	0.1023
Married with sp	0.3749	0.3417		0.5646	0.8063
Married no sp	0.1179	0.1273		0.0647	0.0475
Single	0.4907	0.5128		0.3643	0.1462
European	0.0956	0.0778	0.0200	0.1974	0.1427
Asian	0.2717	0.2542	0.1200	0.3717	0.3282

Table 14 (continued)

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
American	0.2311	0.2412	0.2400	0.1733	0.2548
African	0.0637	0.0391	0.0200	0.2045	0.1000
Mexican	0.3379	0.3877	0.5900	0.0531	0.1743
Sex	0.4626	0.4469	0.4400	0.5525	0.5140
California	0.2046	0.2096	0.2400	0.1762	0.0000
Texas	0.1021	0.1068	0.1400	0.0755	0.0000
Florida	0.0904	0.0911	0.0800	0.0859	0.0000
Arizona	0.0308	0.0335	0.0500	0.0156	0.0000
New York	0.0966	0.0933	0.0500	0.1155	0.0000

Appendix 3: Different samples

This appendix shows the distribution of legal and illegal immigrants using different samples of immigrants from the ACS. We use samples of immigrants from the ACS 2007 who migrated after 2000, after 1990, or after 1980.

Table 15 Legal and illegal distributions from ACS 2007—with immigrants since 1980

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0717	0.0718		0.0716	0.0845
Elementary	0.1558	0.1691		0.1051	0.1277
Junior	0.1341	0.1362		0.1261	0.1517
High school	0.2921	0.2929		0.2891	0.2705
Some college	0.2484	0.2291		0.3217	0.2634
Higher education	0.0979	0.1009		0.0865	0.1023
Married with sp	0.4307	0.3975		0.5571	0.8063
Married no sp	0.1489	0.1584		0.1127	0.0475
Single	0.4196	0.4431		0.3300	0.1462
European	0.0926	0.0678	0.0200	0.1870	0.1427
Asian	0.2528	0.2163	0.1200	0.3915	0.3282
American	0.2386	0.2470	0.2400	0.2065	0.2548
African	0.0513	0.0282	0.0200	0.1394	0.1000
Mexican	0.3647	0.4406	0.5900	0.0757	0.1743
Sex	0.4667	0.4407	0.4400	0.5657	0.5140
California	0.2128	0.2145	0.2400	0.2064	0.0000

Table 15 (continued)

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Texas	0.1038	0.1115	0.1400	0.0746	0.0000
Florida	0.0974	0.0953	0.0800	0.1052	0.0000
Arizona	0.0328	0.0364	0.0500	0.0192	0.0000
New York	0.0944	0.0889	0.0500	0.1156	0.0000

Data: ACS 2007, excluding students, Canadians and Australians, and including professionals. Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer et al. (2008). NIS = statistics on legal green card holders from the 2003 NIS (also used in computing the weights). Estimation using the legal/illegal probability weights from ACS 2007, Table 3 (sampling weights rescaled to account for differences in nonresponse rates). For a description of education categories, see footnotes to Table 3

Table 16 Legal and illegal distributions from ACS 2007—with immigrants since 1990

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0661	0.0658		0.0671	0.0845
Elementary	0.1541	0.1657		0.1048	0.1277
Junior	0.1395	0.1412		0.1324	0.1517
High school	0.2948	0.2956		0.2912	0.2705
Some college	0.2482	0.2319		0.3170	0.2634
Higher education	0.0974	0.0998		0.0874	0.1023
Married with sp	0.5157	0.4931		0.6114	0.8063
Married no sp	0.1098	0.1168		0.0802	0.0475
Single	0.3743	0.3899		0.3084	0.1462
European	0.1086	0.0801	0.0200	0.2293	0.1427
Asian	0.2553	0.2230	0.1200	0.3917	0.3282
American	0.2280	0.2361	0.2400	0.1940	0.2548
African	0.0418	0.0241	0.0200	0.1168	0.1000
Mexican	0.3663	0.4367	0.5900	0.0682	0.1743
Sex	0.4779	0.4572	0.4400	0.5659	0.5140
California	0.2278	0.2307	0.2400	0.2151	0.0000
Texas	0.1090	0.1189	0.1400	0.0671	0.0000
Florida	0.0893	0.0871	0.0800	0.0983	0.0000
Arizona	0.0285	0.0317	0.0500	0.0151	0.0000
New York	0.1041	0.0969	0.0500	0.1346	0.0000

Data: ACS 2007, excluding students, Canadians and Australians, and including professionals. Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer et al. (2008). NIS = statistics on legal green card holders from the 2003 NIS (also used in computing the weights). Estimation using the legal/illegal probability weights from ACS 2007, Table 3 (sampling weights rescaled to account for differences in nonresponse rates). For a description of education categories, see footnotes to Table 3

Table 17 Legal and illegal distributions from ACS 2007—with immigrants since 2000

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0717	0.0718		0.0716	0.0845
Elementary	0.1558	0.1691		0.1051	0.1277
Junior	0.1341	0.1362		0.1261	0.1517
High school	0.2921	0.2929		0.2891	0.2705
Some college	0.2484	0.2291		0.3217	0.2634
Higher education	0.0979	0.1009		0.0865	0.1023
Married with sp	0.4307	0.3975		0.5571	0.8063
Married no sp	0.1489	0.1584		0.1127	0.0475
Single	0.4196	0.4431		0.3300	0.1462
European	0.0926	0.0678	0.0200	0.1870	0.1427
Asian	0.2528	0.2163	0.1200	0.3915	0.3282
American	0.2386	0.2470	0.2400	0.2065	0.2548
African	0.0513	0.0282	0.0200	0.1394	0.1000
Mexican	0.3647	0.4406	0.5900	0.0757	0.1743
Sex	0.4667	0.4407	0.4400	0.5657	0.5140
California	0.2128	0.2145	0.2400	0.2064	0.0000
Texas	0.1038	0.1115	0.1400	0.0746	0.0000
Florida	0.0974	0.0953	0.0800	0.1052	0.0000
Arizona	0.0328	0.0364	0.0500	0.0192	0.0000
New York	0.0944	0.0889	0.0500	0.1156	0.0000

$^{\text{Data:}}$ ACS 2007, excluding students, Canadians and Australians, and including professionals. Benchmarks: $^{\text{DHS}}$= estimation on illegal demographics by the Department of Homeland Security; see Hoefer et al. (2008)

$^{\text{NIS}}$= statistics on legal green card holders from the 2003 NIS (also used in computing the weights). Estimation using the legal/illegal probability weights from ACS 2007, Table 3 (sampling weights rescaled to account for differences in nonresponse rates). For a description of education categories, see footnotes to Table 3

Appendix 4: Sensitivity on different response rates

The following tables present some sensitivity changing the assumptions on nonresponse rates. The first two tables resume all the estimation results, and the following tables use the results from the estimations to create the weights for legal and illegal and compute the means for several characteristics. The assumptions for models M1 to M6 (M6 is the correction used in the main text) for nonresponse rates of illegal and legal immigrants are as follows:

$$\begin{aligned}
\text{M1} &= \text{15 and 10 \%}\\
\text{M2} &= \text{15 and 5 \%}\\
\text{M3} &= \text{15 and 2.5 \%}\\
\text{M4} &= \text{10 and 10 \%}\\
\text{M5} &= \text{10 and 5 \%}\\
\text{M6} &= \text{10 and 2.5 \%}
\end{aligned}$$

Table 18 Probit results: conditional probability of being a legal immigrant from NIS and ACS 2007

	M1	M2	M3	M4	M5	M6
Constant	−2.4475	−2.4690	−2.4793	−2.4244	−2.4463	−2.4567
	(0.1744)	(0.1717)	(0.1704)	(0.1773)	(0.1745)	(0.1732)
Female	0.1203	0.1192	0.1186	0.1215	0.1204	0.1198
	(0.0482)	(0.0475)	(0.0472)	(0.0489)	(0.0482)	(0.0479)
Age	0.0510	0.0505	0.0503	0.0515	0.0510	0.0508
	(0.0041)	(0.0040)	(0.0040)	(0.0041)	(0.0041)	(0.0041)
Elementary	0.6431	0.6374	0.6347	0.6491	0.6434	0.6407
	(0.1573)	(0.1551)	(0.1540)	(0.1597)	(0.1574)	(0.1564)
Junior high	1.0860	1.0770	1.0727	1.0955	1.0865	1.0822
	(0.1703)	(0.1677)	(0.1665)	(0.1731)	(0.1705)	(0.1692)
High school	0.5468	0.5420	0.5397	0.5520	0.5471	0.5448
	(0.1331)	(0.1312)	(0.1303)	(0.1351)	(0.1332)	(0.1323)
College degree	0.3419	0.3390	0.3375	0.3450	0.3421	0.3406
	(0.1298)	(0.1279)	(0.1270)	(0.1319)	(0.1299)	(0.1290)
Higher education	−0.0981	−0.0973	−0.0969	−0.0989	−0.0981	−0.0978
	(0.1422)	(0.1401)	(0.1392)	(0.1444)	(0.1423)	(0.1413)
Married spouse present	0.1466	0.1456	0.1451	0.1477	0.1467	0.1462
	(0.0546)	(0.0539)	(0.0535)	(0.0554)	(0.0547)	(0.0543)
Married spouse not present	−0.5196	−0.5150	−0.5128	−0.5245	−0.5199	−0.5176
	(0.0856)	(0.0847)	(0.0843)	(0.0866)	(0.0857)	(0.0852)
Mex*¡elementary	−0.7472	−0.7402	−0.7369	−0.7546	−0.7476	−0.7442
	(0.1916)	(0.1888)	(0.1874)	(0.1948)	(0.1918)	(0.1904)
Mex*elementary	−1.7050	−1.6896	−1.6822	−1.7213	−1.7059	−1.6985
	(0.1793)	(0.1773)	(0.1764)	(0.1815)	(0.1794)	(0.1785)
Mex*junior high	−2.1413	−2.1228	−2.1139	−2.1610	−2.1424	−2.1335
	(0.2059)	(0.2038)	(0.2028)	(0.2083)	(0.2060)	(0.2050)
Mex*high school	−1.6497	−1.6350	−1.6279	−1.6654	−1.6505	−1.6434
	(0.1526)	(0.1512)	(0.1505)	(0.1542)	(0.1527)	(0.1520)
Mex*college degree	−1.5952	−1.5811	−1.5744	−1.6102	−1.5960	−1.5892
	(0.2109)	(0.2099)	(0.2093)	(0.2121)	(0.2110)	(0.2105)
Mex*higher education	−1.0125	−1.0032	−0.9987	−1.0224	−1.0130	−1.0085
	(0.3872)	(0.3854)	(0.3846)	(0.3891)	(0.3874)	(0.3865)

Table 18 (continued)

	M1	M2	M3	M4	M5	M6
America	−0.8563	−0.8483	−0.8445	−0.8648	−0.8567	−0.8529
	(0.1076)	(0.1058)	(0.1050)	(0.1095)	(0.1077)	(0.1068)
Africa	0.5317	0.5278	0.5259	0.5357	0.5319	0.5300
	(0.1411)	(0.1376)	(0.1359)	(0.1450)	(0.1413)	(0.1396)
Asia	−0.3593	−0.3556	−0.3538	−0.3633	−0.3595	−0.3577
	(0.0921)	(0.0902)	(0.0894)	(0.0942)	(0.0922)	(0.0913)
q	0.4010	0.3857	0.3785	0.4175	0.4019	0.3945
	(0.0474)	(0.0472)	(0.0472)	(0.0476)	(0.0474)	(0.0473)
LogLik	−3,574.22	−3,574.23	−3,574.23	−3,574.21	−3,574.22	−3,574.22
No. of obs	9,464	9,464	9,464	9,464	9,464	9,464

Table 19 Legal and illegal distributions from ACS 2007 using immigrant weights from 2007 ACS M1

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0610	0.0674		0.0375	0.0845
Elementary	0.1606	0.1821		0.0816	0.1277
Junior	0.1399	0.1437		0.1258	0.1517
High school	0.2943	0.2931		0.2988	0.2705
Some college	0.2453	0.2160		0.3526	0.2634
Higher education	0.0989	0.0977		0.1037	0.1023
Married with sp	0.4348	0.3841		0.6208	0.8063
Married no sp	0.1427	0.1605		0.0776	0.0475
Single	0.4221	0.4551		0.3015	0.1462
European	0.0892	0.0619	0.0200	0.1894	0.1427
Asian	0.2315	0.1947	0.1200	0.3664	0.3282
American	0.2321	0.2378	0.2400	0.2111	0.2548
African	0.0483	0.0218	0.0200	0.1455	0.1000
Mexican	0.3989	0.4839	0.5900	0.0876	0.1743
Sex	0.4514	0.4255	0.4400	0.5459	0.5140
California	0.2038	0.2105	0.2400	0.1793	0.0000
Texas	0.1094	0.1181	0.1400	0.0777	0.0000
Florida	0.0911	0.0901	0.0800	0.0949	0.0000
Arizona	0.0330	0.0372	0.0500	0.0175	0.0000
New York	0.0890	0.0821	0.0500	0.1147	0.0000

Table 20 Legal and illegal distributions from ACS 2007 using immigrant weights from 2007 ACS M2

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0610	0.0672		0.0372	0.0845
Elementary	0.1606	0.1813		0.0812	0.1277
Junior	0.1399	0.1434		0.1264	0.1517
High school	0.2943	0.2930		0.2993	0.2705
Some college	0.2453	0.2172		0.3529	0.2634
Higher education	0.0989	0.0979		0.1030	0.1023
Married with sp	0.4348	0.3860		0.6223	0.8063
Married no sp	0.1427	0.1598		0.0771	0.0475
Single	0.4221	0.4539		0.3004	0.1462
European	0.0892	0.0629	0.0200	0.1902	0.1427
Asian	0.2315	0.1964	0.1200	0.3665	0.3282
American	0.2321	0.2379	0.2400	0.2097	0.2548
African	0.0483	0.0225	0.0200	0.1475	0.1000
Mexican	0.3989	0.4804	0.5900	0.0861	0.1743
Sex	0.4514	0.4266	0.4400	0.5464	0.5140
California	0.2038	0.2102	0.2400	0.1792	0.0000
Texas	0.1094	0.1177	0.1400	0.0776	0.0000
Florida	0.0911	0.0902	0.0800	0.0947	0.0000
Arizona	0.0330	0.0370	0.0500	0.0174	0.0000
New York	0.0890	0.0824	0.0500	0.1147	0.0000

Table 21 Legal and illegal distributions from ACS 2007 using immigrant weights from 2007 ACS M3

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0610	0.0671		0.0371	0.0845
Elementary	0.1606	0.1809		0.0810	0.1277
Junior	0.1399	0.1432		0.1267	0.1517
High school	0.2943	0.2930		0.2996	0.2705
Some college	0.2453	0.2178		0.3530	0.2634
Higher education	0.0989	0.0980		0.1027	0.1023
Married with sp	0.4348	0.3869		0.6230	0.8063
Married no sp	0.1427	0.1594		0.0769	0.0475
Single	0.4221	0.4533		0.2999	0.1462
European	0.0892	0.0634	0.0200	0.1907	0.1427
Asian	0.2315	0.1972	0.1200	0.3665	0.3282
American	0.2321	0.2379	0.2400	0.2089	0.2548
African	0.0483	0.0228	0.0200	0.1485	0.1000
Mexican	0.3989	0.4787	0.5900	0.0854	0.1743
Sex	0.4514	0.4271	0.4400	0.5466	0.5140
California	0.2038	0.2101	0.2400	0.1791	0.0000
Texas	0.1094	0.1176	0.1400	0.0775	0.0000
Florida	0.0911	0.0902	0.0800	0.0946	0.0000
Arizona	0.0330	0.0369	0.0500	0.0173	0.0000
New York	0.0890	0.0825	0.0500	0.1148	0.0000

Table 22 Legal and illegal distributions from ACS 2007 using immigrant weights from 2007 ACS M4

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0610	0.0676		0.0378	0.0845
Elementary	0.1606	0.1831		0.0821	0.1277
Junior	0.1399	0.1441		0.1251	0.1517
High school	0.2943	0.2932		0.2983	0.2705
Some college	0.2453	0.2146		0.3524	0.2634
Higher education	0.0989	0.0974		0.1043	0.1023
Married with sp	0.4348	0.3820		0.6191	0.8063
Married no sp	0.1427	0.1612		0.0780	0.0475
Single	0.4221	0.4564		0.3028	0.1462
European	0.0892	0.0607	0.0200	0.1884	0.1427
Asian	0.2315	0.1929	0.1200	0.3662	0.3282
American	0.2321	0.2376	0.2400	0.2127	0.2548
African	0.0483	0.0211	0.0200	0.1434	0.1000
Mexican	0.3989	0.4877	0.5900	0.0892	0.1743
Sex	0.4514	0.4244	0.4400	0.5453	0.5140
California	0.2038	0.2108	0.2400	0.1795	0.0000
Texas	0.1094	0.1185	0.1400	0.0779	0.0000
Florida	0.0911	0.0899	0.0800	0.0952	0.0000
Arizona	0.0330	0.0374	0.0500	0.0175	0.0000
New York	0.0890	0.0817	0.0500	0.1146	0.0000

Table 23 Legal and illegal distributions from ACS 2007 using immigrant weights from 2007 ACS M5

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below elementary	0.0610	0.0674		0.0375	0.0845
Elementary	0.1606	0.1822		0.0817	0.1277
Junior	0.1399	0.1437		0.1257	0.1517
High school	0.2943	0.2931		0.2988	0.2705
Some college	0.2453	0.2159		0.3526	0.2634
Higher education	0.0989	0.0977		0.1037	0.1023
Married with sp	0.4348	0.3840		0.6207	0.8063
Married no sp	0.1427	0.1605		0.0776	0.0475
Single	0.4221	0.4551		0.3016	0.1462
European	0.0892	0.0618	0.0200	0.1893	0.1427
Asian	0.2315	0.1946	0.1200	0.3664	0.3282
American	0.2321	0.2378	0.2400	0.2112	0.2548
African	0.0483	0.0217	0.0200	0.1454	0.1000
Mexican	0.3989	0.4841	0.5900	0.0877	0.1743
Sex	0.4514	0.4255	0.4400	0.5459	0.5140
California	0.2038	0.2105	0.2400	0.1793	0.0000
Texas	0.1094	0.1181	0.1400	0.0777	0.0000
Florida	0.0911	0.0900	0.0800	0.0949	0.0000
Arizona	0.0330	0.0372	0.0500	0.0175	0.0000
New York	0.0890	0.0820	0.0500	0.1147	0.0000

References

Burtless G, Singer A (2011) The earnings and social security contributions of documented and undocumented Mexican immigrants, No. 2 in Working Paper. Boston College Retirement Research Center

Camarota S, Jeffrey C (2004) Assessing the quality of data collected on the foreign born: an evaluation of the american comm unity survey (ACS). Methodology and data quality. COPAFS (The Council of Professional Associations on Federal Statistics)

Durand J, Massey D (2006) What we learned from the Mexican migration Project vol. Crossing the border: research from the Mexican migration project. Russell Sage Foundation, New York

Fortin N, Lemieux T, Firpo S (2011) Decomposition methods in economics Handbook of labor economics, chap. 1, vol 4. Elsevier, New York, pp 1–102

Hoefer M, Rytina N, Baker BC (2008) Estimates of the unauthorized immigrant population residing in the United States: January 2007, Population Estimates. U.S. Department of Homeland Security, Office of Immigration Statistics

Lancaster T, Imbens G (1996) Case-control studies with contaminated controls. J Econ 71(1–2):145–160

Passel J (2006) The size and characteristics of the unauthorized migrant population in the U.S. estimates based on the March 2005 current population survey. Research Report, PEW Hispanic Center

Passel J, Randolph C, Fix M (2004) Undocumented Immigrants: facts and figures. Immigration studies program. Urban Institute, Washington DC

Ridder G, Moffitt R (2007) The econometrics of data combination, chapter 75. Part 2 of Handbook of econometrics, vol 6. Elsevier, New York, pp 5469–5547

Rosenblum M (2012) Border security: immigration enforcement between ports of entry. No. 2 in congressional research service. Washington DC