NBER WORKING PAPER SERIES

ATTENUATION BIAS IN MEASURING THE WAGE IMPACT OF IMMIGRATION

Abdurrahman Aydemir
George J. Borjas

Working Paper 16229
http://www.nber.org/papers/w16229

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2010

We are grateful to Joshua Angrist, Sue Dynarski, Richard Freeman, Daniel Hamermesh, Larry Katz, Robert Moffitt, Jeffrey Smith, Douglas Staiger, and especially to Alberto Abadie for helpful comments and suggestions on earlier drafts of this paper. Most of the work on this article was completed while Dr. Aydemir was employed at Statistics Canada in Ottawa, Canada. The authors are grateful to Statistics Canada for their invaluable research support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Attenuation Bias in Measuring the Wage Impact of Immigration
Abdurrahman Aydemir and George J. Borjas
NBER Working Paper No. 16229
July 2010
JEL No. C1,J0,J6

ABSTRACT

Although economic theory predicts an inverse relation between relative wages and immigration-induced supply shifts, it has been difficult to document such effects. The weak evidence may be partly due to sampling error in a commonly used measure of the supply shift, the immigrant share of the workforce. After controlling for permanent factors that determine wages in specific labor markets, little variation remains in the immigrant share. We find significant sampling error in this measure of supply shifts in Canadian and U.S. Census data. Correcting for the resulting attenuation bias can substantially increase existing estimates of the wage impact of immigration.

Abdurrahman Aydemir
Sabanci University
Faculty of Arts and Social Sciences
Istanbul, Turkey
aaydemir@sabanciuniv.edu

George J. Borjas
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138
and NBER
gborjas@harvard.edu

Attenuation Bias in Measuring the Wage Impact of Immigration

Abdurrahman Aydemir and George J. Borjas*

I. Introduction

The textbook model of a competitive labor market has clear and unambiguous implications about how wages should adjust to an immigration-induced labor supply shift. In particular, higher levels of immigration should lower the wage of competing workers, at least in the short run.

Despite the common-sense intuition behind this prediction, the economics literature has found it difficult to document the inverse relation between wages and immigration-induced supply shifts. Much of the literature estimates the labor market impact of immigration in a receiving country by comparing economic conditions across local labor markets in that country. Although there is a great deal of dispersion in the measured impact across these geographic labor market studies, the estimates tend to cluster around zero. This finding has been interpreted as indicating that immigration has little impact on the receiving country's wage structure.[1]

One problem with this interpretation is that the "spatial correlation"—the correlation between labor market outcomes and immigration across local labor markets—may not truly capture the wage impact of immigration if native workers (or capital) respond by moving their

---

* Dr. Aydemir is an Assistant Professor of Economics, Sabancı University, İstanbul, Turkey; Dr. Borjas is a Professor of Economics and Social Policy at the Harvard Kennedy School, Cambridge, MA, and a Research Associate at the National Bureau of Economic Research. We are grateful to Joshua Angrist, Sue Dynarski, Richard Freeman, Daniel Hamermesh, Larry Katz, Robert Moffitt, Jeffrey Smith, Douglas Staiger, and especially to Alberto Abadie for helpful comments and suggestions on earlier drafts of this paper. Most of the work on this article was completed while Dr. Aydemir was employed at Statistics Canada in Ottawa, Canada. The authors are grateful to Statistics Canada for their invaluable research support.

[1] Representative studies include Altonji and Card (1991), Borjas (1987), Borjas, Freeman, and Katz (1997), Card (1991, 2001), Grossman (1982), Hartog and Zorlu (2005), LaLonde and Topel (1991), Pischke and Velling (1997), and Schoeni (1997). Friedberg and Hunt (1995), Smith and Edmonston (1997), and Longhi et al. (2005) survey the literature.

inputs to localities seemingly less affected by the immigrant supply shock.[2] Because these flows arbitrage regional wage differences, the wage impact of immigration may perhaps be best measured at the national level. Borjas (2003) used this insight to examine if the evolution of wages in particular skill groups—defined in terms of both educational attainment and years of work experience—were related to the immigrant supply shocks affecting those groups. In contrast to the geographic labor market studies, the national labor market evidence indicated that wage growth was strongly and inversely related to immigration-induced supply increases.[3]

A number of papers have replicated the national-level approach, with mixed results. These initial replications, therefore, seem to suggest that the national labor market approach may find itself with as many different types of results as the spatial correlation approach that it conceptually and empirically attempted to replace. For example, Mishra (2007) applies the framework to the Mexican labor market and finds significant positive wage effects of emigration on wages in Mexico. On the other hand, Bonin (2005) applies the framework to the German labor market and reports a very weak impact of supply shifts on the wage structure. Aydemir and Borjas (2007) apply the approach to both Canadian and Mexican Census data and find a strong inverse relation between wages and immigration-induced supply shifts. In contrast, Bohn and

---

[2] The literature has not reached a consensus on whether native workers respond to immigration by voting with their feet and moving to other areas. Filer (1992), Frey (1995), and Borjas (2006) find a strong internal migration response, while Card (2001) and Kritz and Gurak (2001) find little connection between native migration and immigration. It is worth noting that the spatial correlation will also be positively biased if income-maximizing immigrants choose to locate in high-wage areas, creating a spurious correlation between immigrant supply shocks and wages. Borjas (2001) and Cadena (2010) show that immigrants tend to settle in those cities that offer the best economic opportunities for the skills they have to offer. Alternative modes of market adjustment are studied by Lewis (2005), who examines the link between immigration and the input mix used by firms, and Saiz (2003), who examines how rental prices adjusted to the Mariel immigrant influx.

[3] Note that the classification of workers into narrowly defined skill groups (based on education and experience) represents an empirical strategy that should also be pursued by the geographic labor market studies—after all, both the national and geographic labor market studies attempt to estimate the impact of an immigration-induced increase in the number of workers with a particular set of skills on the wage of comparable pre-existing

Sanders (2005) use publicly available Canadian data and report weak wage effects in the Canadian labor market.

This paper argues that the differences in estimated coefficients across the fast-growing set of national labor market studies, as well as many of the very weak coefficients reported in the spatial correlation literature, may well be explained by a simple statistical fact: There is a lot of sampling error in the measures of the immigrant supply shift commonly used in the literature, and this sampling error leads to substantial attenuation bias in the estimated wage impact of immigration.[4]

Measurement error plays a central role in these studies because of the longitudinal nature of the exercise that is conducted. The immigration-induced supply shift is often measured by the "immigrant share," the fraction of the workforce in a particular labor market that is foreign-born. The analyst then examines the relation between the wage and the immigrant share *within* a particular labor market. To net out market-specific wage effects, the study typically includes various vectors of fixed effects (e.g., regional fixed effects or skill-level fixed effects) that absorb these permanent factors. The inclusion of these fixed effects implies that there is very little identifying variation left in the variable that captures the immigrant supply shift, permitting any sampling error in the immigrant share to play a disproportionately large role. As a result, even very small amounts of sampling error get magnified and easily dominate the remaining variation in the immigrant share.

---

workers. As will be discussed below, however, many of the geography-based studies ignore the skill composition of the immigrant workforce when estimating the wage impact of immigration.

[4] The biases resulting from sampling error are well known to be important in empirical work outside the immigration context. For example, Paxson and Waldfogel (2002) note that corrections for sampling error have very large effects; in some cases more than doubling parameter estimates when they investigate the impact of parental economic circumstances on child maltreatment using state-level panel data.

Because the immigrant share variable is a proportion, its sampling error can be easily derived from the properties of the hypergeometric distribution. The statistical properties of this random variable provide a great deal of information that can be used to measure the extent of attenuation bias in these types of models as well as to construct relatively simple corrections for measurement error.

Our empirical analysis uses data for both Canada and the United States to show the numerical importance of sampling error in attenuating the wage impact of immigration. We have access to the *entire* Census files maintained by Statistics Canada. These Census files represent a sizable sampling of the Canadian population: a 33.3 percent sample in 1971 and a 20 percent sample thereafter.[5] The application of the national labor market model proposed by Borjas (2003) to these entire samples reveals a significant negative correlation between wages of specific skill groups and immigrant supply shifts. It turns out, however, that when the *identical* regression is estimated in smaller samples (even on those that are publicly released by Statistics Canada), the regression coefficient is numerically much smaller and much less likely to be statistically significant. We also find the same pattern of attenuation bias in our study of U.S. Census data. A regression model estimated on the largest samples available reveals significant effects, but the effects become exponentially weaker as the analyst calculates the immigrant share on progressively smaller samples.

[5] These confidential files are the largest available micro data files in Canada that provide information on citizenship, immigration, schooling, labor market activities, and earnings. The confidential data is available on a cost-recovery basis to researchers not employed by Statistics Canada who abide by the agency's confidentiality rules.

II. Framework

We are interested in estimating the wage impact of immigration by looking at wage variation across labor markets. A labor market k ($k = 1,\ldots,K$) may be defined in terms of skills, geographic regions, and/or time. The available data has been aggregated to the level of the labor market and typically reports the wage level and the size of the immigrant supply shock in each market. The generic regression model estimated in much of the literature can be summarized as:

$$w_k = \beta\pi_k + \sum_h \alpha_h z_{kh} + \varepsilon_k , \qquad (1)$$

where w_k gives the log wage in labor market k; π_k gives the immigrant share in the labor market (i.e., the fraction of the workforce that is foreign-born); the variables in the vector Z $(h = 1,\ldots,H)$ are control variables that may include period fixed effects, region fixed effects, skill fixed effects, and any other variables that generate differences in wage levels across labor markets; and ε is an i.i.d. error term, with mean 0 and variance σ_ε^2.

A crucial characteristic of this type of empirical exercise is that the analyst typically calculates the immigrant share from the microdata available for labor market k. This type of calculation introduces sampling error in the key independent variable in equation (1), and introduces the possibility that the coefficient β may be inconsistently estimated.[6]

To fix ideas, suppose that all other variables in the regression model are measured correctly. Suppose further that the only type of measurement error in the observed immigrant share p_k is the one that arises due to sampling error and not to any possible misclassification of

[6] Note that while we use the terms measurement error and sampling error interchangeably in the rest of the text, we will mainly be concerned with errors that arise due to sampling.

workers by immigrant status.[7] The relation between the observed immigrant share and the true immigrant share in the labor market is given by:

$$p_k = \pi_k + u_k \,. \tag{2}$$

When a data sample of size n_k is obtained by sampling with replacement from a population of size N_k, the observed immigrant share is the mean of a sample of independent Bernoulli draws, so that $E(u_k) = 0$ and $Var(u_k) = \pi_k(1 - \pi_k)/n_k$. Census sampling, however, is without replacement and the error term in (2) has a hypergeometric distribution with $E(u_k) = 0$ and

$$Var(u_k) = \frac{\pi_k(1-\pi_k)}{n_k} \times \frac{N_k - n_k}{N_k - 1}.$$

The size of the population in the labor market, N_k, is not typically observed, but the expected value of the ratio n_k/N_k is known and is simply the sampling rate (τ) that generates the Census sample (e.g., a 1/1000 sample). We approximate the variance of the error term in (2) by $Var(u_k) \approx (1 - \tau)\, \pi_k (1 - \pi_k)/n_k$. Note that the variance of the sampling error has a simple binomial structure for very small sampling rates.[8] Further, u_k and π_k are mean-independent, implying $Cov(\pi_k, u_k) = 0$. We will show below that these statistical properties of the sampling error have important implications for the size of the attenuation bias in estimates of the wage impact of

---

[7] It is likely that the results reported in many studies (particularly those conducted in the 1980s and early 1990s) are contaminated by a different type of measurement error. In particular, these studies often examined the impact of immigrant supply shocks on the wage of particular skill groups, such as high school dropouts. However, the measure of the immigrant supply shock used in these studies often ignored the skill composition of the foreign-born workforce and was simply defined as the immigrant share in the labor market (see, for example, Altonji and Card, 1991; Borjas, 1987; and LaLonde and Topel, 1991).

[8] Conversely, for very large sampling rates the sample approximates the population and there is little sampling error in the observed measure of the immigrant share.

immigration. They also provide relatively simple ways for correcting the estimates for the impact of sampling error.

The easiest way of quantifying the magnitude of the bias in this context is to follow the standard method in the measurement error literature, and simply examine the asymptotic properties of the OLS estimator as sample size *K* goes to infinity. In particular, it is well known that the probability limit of $\hat{\beta}$ in a multivariate regression model when only the regressor p_k is measured with error is:[9]

$$\text{plim}\ \hat{\beta} = \beta\left(1 - \frac{\text{plim}\frac{1}{K}\sum_k u_k^2}{(1-R^2)\sigma_p^2}\right), \tag{3}$$

where σ_p^2 is the variance of the observed immigrant share across the *K* labor markets, and R^2 is the multiple correlation of an auxiliary regression that relates the observed immigrant share to all other right-hand-side variables in the model. The term $(1-R^2)\sigma_p^2$, therefore, gives the "purged" variance, the variance of the observed immigrant share that remains unexplained after controlling for all other variables in the regression model.

As noted above, the typical study in the literature pools data on particular labor markets over time and adds fixed effects that net out persistent wage differences across labor markets as well as period effects. This type of regression model, of course, is effectively differencing the

---

[9] Maddala (1992, pp. 451-454) presents a particularly simple derivation of equation (3) when the regression has two explanatory variables; see also Cameron and Trivedi (2005, p. 904). Garber and Keppler (1980), and Levi (1973). Griliches, and Hausman (1986), Bound and Krueger (1991), and McKinnish (2008) discuss measurement error in panel data models and Bound, Brown and Mathiowetz (2001) provide an excellent survey of the measurement error literature.

data so that the wage impact of immigration is identified from within-market changes in the immigrant share. The multiple correlation of the auxiliary regression in this type of longitudinal study will typically be very high, usually above 0.9. As a result, much of the systematic variation in the immigrant share is "explained away," and the measurement error introduced by the sampling error plays a disproportionately large role in the estimation.

One problem with this approach is that the asymptotic exercise involves letting the number of markets K go to infinity while holding fixed the sample size n_k used to estimate the immigrant share in the market. An alternative (and perhaps more sensible) exercise would be to let both the number of markets K and the sample size n_k go to infinity. In this case, of course, the estimator would be consistent as there would not be any sampling error. It would be of interest to derive the finite-sample properties for the OLS estimator $\hat{\beta}$—in the sense that the number of markets K is "small." Although finite-sample properties of regression coefficients in measurement error models are difficult to derive explicitly, the results presented in Richardson and Wu (1970, p. 729) for the classical model suggest that even when K is of moderate size (around 50 or 100) the expected value of the OLS coefficient $\hat{\beta}$ can be closely approximated by the asymptotic formula in (3).[10] Hence equation (3) may be a potentially valuable approximation of the impact of attenuation bias in the immigration context.

---

[10] Richardson and Wu (1970) examine the finite-sample properties of the coefficients in a bivariate regression model where both the dependent and the independent variables have classical measurement error. Our model differs mainly in that it also includes a vector Z of correctly measured regressors. We can reinterpret our multivariate regression model as a bivariate regression where the wage is being regressed on the purged residual of the immigrant share. The Richardson-Wu results may be applicable if we could interpret our observed purged residual (i.e., the residual from the auxiliary regression of the observed immigrant share on Z, with associated coefficient vector ϕ') as the sum of the "true" purged residual (i.e., the residual from the unfeasible auxiliary regression of the true immigrant share on Z, with coefficient vector ϕ'') and the measurement error. It is easy to show that the observed purged residual equals the "true" purged residual plus the measurement error u plus a term involving the difference $(\phi' - \phi'')$. This difference has expected value of zero since errors in the dependent variable do not lead to bias. We conducted a number of Monte Carlo simulations that indicated that this term accounts for

In the Mathematical Appendix, we show that the probability limit of the average of the square of the error terms as $K \to \infty$ in (3) is:

$$\text{plim}\frac{1}{K}\sum_k u_k^2 = (1-\tau)E\left(\frac{\pi_k(1-\pi_k)}{n_k}\right), \tag{4}$$

where the expectation in (4) is taken across the K labor markets. Note that the average of the squared error terms goes to zero if the sample size n_k also goes to infinity. Combining results, we can then write:

$$\text{plim}\hat{\beta} = \beta\left(1-(1-\tau)\frac{E[\pi_k(1-\pi_k)/n_k]}{(1-R^2)\sigma_p^2}\right). \tag{5}$$

Equation (5) imposes an important restriction on the magnitude of the sampling error. The expected sampling error given by (4) must be less than the unexplained portion of the variance in the immigrant share (in other words, the variance due to sampling error cannot be larger than the variance that remains after controlling for other observable characteristics). This restriction implies that in situations where sampling error tends to be large and where there is little variance left in the immigrant share after controlling for variation in the other variables, the classical errors-in-variables model may be uninformative and it may be impossible to retrieve

---

only 0.2 percent of the variance in the observed purged residual when $K = 50$, and less than 0.02 percent when $K = 400$. Hence the Richardson-Wu results may provide a reasonably accurate approximation of the finite-sample properties of the OLS estimator $\hat{\beta}$.

information about the value of the true parameter from observed data. This restriction is often violated when the immigrant share is calculated in relatively small samples.

The violations may arise for two reasons. First, any calculation of the expected sampling error in (5) requires that we approximate the true immigrant share π_k with the observed immigrant share p_k. This approximation introduces errors, making it possible for the estimate of the expected sampling error to exceed the adjusted variance in small samples.

Second, we assumed that the only source of measurement error in the observed immigrant share is sampling error. There could well be other types of errors, such as classification errors of immigrant status (Aigner, 1973; Freeman, 1981; and Kane, Rouse, and Staiger, 1999). In relatively small samples, where the sampling error already accounts for a very large fraction of the adjusted variance, even a minor misclassification problem could easily lead to a violation of the restriction implied by equation (5).

It is useful to present an approximation to equation (5) that gives a back-of-the-envelope formula for estimating the importance of attenuation bias. In particular, suppose that we calculate the average sampling error so that larger cells count more than smaller cells. Define the weight $\lambda_k = n_k/n_T$, where n_T gives the total sample size across all K labor markets. We can then rewrite the expectation in (4) as:

$$\begin{aligned} E\left(\frac{\pi_k(1-\pi_k)}{n_k}\right) &= \sum_k \lambda_k \frac{\pi_k(1-\pi_k)}{n_k} \\ &= \sum_k \frac{n_k}{n_T}\frac{\pi_k(1-\pi_k)}{n_k} \\ &= \frac{E[\pi_k(1-\pi_k)]}{\bar{n}}, \end{aligned} \tag{6}$$

where $\bar{n}$ (= n_T/K) is the average per-cell number of observations used to calculate the immigrant share in the various labor markets. It is easy to show that $E[\pi_k(1 - \pi_k)]$ can be closely approximated by the expression $\bar{p}(1-\bar{p})$, where $\bar{p}$ is the average observed immigrant share across the K labor markets.[11] We can then rewrite equation (5) as:

$$\operatorname{plim}\hat{\beta}=\beta\left(1-(1-\tau)\frac{\bar{p}(1-\bar{p})/\bar{n}}{(1-R^2)\sigma_p^2}\right). \tag{7}$$

Equation (7) implies that the percent bias generated by sampling error is given by:

$$\frac{\operatorname{plim}\hat{\beta}-\beta}{\beta}=(1-\tau)\frac{\bar{p}(1-\bar{p})/\bar{n}}{(1-R^2)\sigma_p^2}. \tag{8}$$

The immigrant share in the United States is around 0.1, and we will show below that the variance in the immigrant share across national labor markets defined on the basis of skills (in particular, schooling and work experience) is approximately 0.004. Finally, the explanatory power of the auxiliary regression of the immigrant share on all the other variables in the model (such as fixed effects for education and experience) is very high, on the order of 0.95. Figure 1 illustrates the predicted size of the bias as a function of the per-cell sample size when the sampling rate τ is small ($\tau \to 0$). It is evident that even when the immigrant share is calculated using 1,000 observations per cell there is a remarkably high level of attenuation in the coefficient β. In

[11] The difference between $\bar{\pi}(1-\bar{\pi})/\bar{n}$ and $E[\pi_k(1-\pi_k)]/\bar{n}$ equals $\sigma_\pi^2/\bar{n}$, where $\bar{\pi}=E(\pi_k)$. The approximation, therefore, is quite good for any reasonable value of $\bar{n}$.

particular, the percent bias is 45 percent when the average cell has 1,000 observations, 60 percent when there are 750 observations, 75 percent when there are 600 observations, and the coefficient is completely driven to zero when there are 450 observations.[12]

The figure also reports the results of a similar calculation with data from the Canadian labor market. In Canada, the immigrant share is around 0.2, and we will show below that the variance in the immigrant share across national labor markets (defined by education and experience) is around 0.005. The R^2 of the auxiliary regression is again around 0.95. The fact that the immigrant share is twice as large in Canada implies that the bias is higher than in the United States—for a given mean cell size. In particular, the percent bias is 64 percent when the average cell has 1,000 observations, 85.3 percent when there are 750 observations, and sampling error completely overwhelms the data when there are fewer than 640 observations. It is also worth noting that the hypergeometric distribution of the sampling error—combined with the fact that the longitudinal nature of the exercise removes much of the identifying variation in the immigrant share—implies a quantitatively meaningful bias even when there are as many as 10,000 observations per cell: the percent bias is then 6.4 percent in Canada and 4.5 percent in the United States.

The "back of the envelope" correction implied by equation (8) is likely to be particularly useful in empirical applications that use proportions as independent variables in regression models. Although practical, the correction depends on various assumptions (such as the number of labor markets being very large) that may not be strictly satisfied by the data. As a result, it is important to examine if alternative methods of correcting for attenuation bias lead to generally similar results as the simpler back-of-the-envelope approach. The OLS estimator for β in

[12] The bias cannot be calculated if the average cell size is less than 450. The implied amount of

equation (1) is biased and inconsistent when the regression uses the observed immigrant share because the sampling error *u* creates a non-zero correlation between the resulting error term and *p*. Consistent estimates may be obtained if either some information about the sampling error is known or if an instrument is found that is correlated with π but uncorrelated with *u*. In the next section, we will introduce two alternative methods—based on an approximation of a moment of the sampling error and on the method of instrumental variables—that can also be used in applied work.

Because many of the recent empirical studies in the literature use the seemingly large Public Use Samples of the U.S. Census (which contain individual observations for a 5 percent sample of the population since 1980), it may seem that the number of observations used to calculate the immigrant share is likely to be far higher than just a few hundred (or even a few thousand), so that the attenuation problem would be relatively minor. It turns out, however, that once the analyst begins to define the "labor market" in ever-narrower terms (e.g., skill groups or occupations within a geographic area), it is quite easy for even these very large 5 percent files to yield relatively small samples for the average cell and the attenuation bias can easily become substantively important.

Finally, our analysis assumes that the immigrant share is the only mismeasured variable in the regression model. Deaton (1985) suggests that there may be non-classical errors because the immigrant share is unlikely to be the only variable that is measured less precisely as the cell size gets smaller. The dependent variable (the mean of the log wage in market *k*) also is measured more imprecisely in smaller samples. In some contexts, Deaton (1985) shows that the sampling error between the dependent and independent variables could be correlated.

---

measurement error would then be larger than the unexplained variance in the immigrant share.

Such a correlation, however, does not exist in our context. To see why, consider the nature of the sampling error in the immigrant share. Suppose we happen to sample "too many" natives in market *k*, underestimating the true immigrant share. What is the impact of this sampling error on the calculated mean (log) earnings of native workers in that market? Each additional native that was over-sampled was drawn at random from the population of natives in market *k*. As a result, the expected value of the earnings of the over-sampled natives equals the average earnings of natives in market *k*, implying that the sampling error in mean log earnings is independent from the sampling error in the immigrant share.

III. Data and Results

We use microdata Census files for both Canada and the United States to illustrate the quantitative importance of attenuation bias in estimating the wage impact of immigration. Our study of the Canadian labor market uses all available files from the Canadian Census (1971, 1981, 1986, 1991, 1996, and 2001). Each of these confidential files, resident at Statistics Canada, represents a 20 percent sample of the Canadian population (except for the 1971 file, which represents a 33.3 percent sample). Statistics Canada provides Public Use Microdata Files (PUMFs) to Canadian post-secondary institutions and to other researchers. The PUMFs use a much smaller sampling rate than the confidential files used in this paper. In particular, the 1971 PUMF comprises a 1.0 percent sample of the Canadian population, the 1981 and 1986 PUMFs comprise a 2.0 percent sample, the 1991 PUMF comprises a 3.0 percent sample, the 1996 PUMF comprises a 2.8 percent sample, and the 2001 PUMF comprises a 2.7 percent sample.

Our study of the U.S. labor market uses the 1960, 1970, 1980, 1990 and 2000 Integrated Public Use Microdata Samples (IPUMS) of the decennial Census. The 1960 file represents a 1

percent sample of the U.S. population, the 1970 file represents a 3 percent sample, and the 1980 through 2000 files represent a 5 percent sample.[13] For expositional convenience, we will refer to the data from these five Censuses as the "5 percent file," even though the 5/100 sampling rate only applies to the data collected since 1980.

We restrict the empirical analysis to men aged 18 to 64 who participate in the civilian labor force. The Data Appendix describes the construction of the sample extracts and variables in detail. Our analysis of the U.S. data uses the convention of defining an immigrant as someone who is either a noncitizen or a naturalized U.S. citizen. In the Canadian context, we define an immigrant as someone who reports being a "landed immigrant" (i.e., a person who has been granted the right to live in Canada permanently by immigration authorities), and is either a noncitizen or a naturalized Canadian citizen.[14]

A. National Labor Market

As noted earlier, Borjas (2003) suggests that the wage impact of immigration can perhaps best be measured by looking at the evolution of wages in the national labor market for different skill groups. He defines skill groups in terms of both educational attainment and work experience to allow for the possibility that workers who belong to the same education groups but differ in their work experience are not perfect substitutes.

---

[13] We created the 3 percent 1970 sample by pooling the 1/100 Form 1 state, metropolitan area, and neighborhood files. These three samples are independent, so that the probability that a particular person appears in more than one of these samples is negligible.

[14] Since 1991, the Canadian Censuses include non-permanent residents. This group includes those residing in Canada on an employment authorization, a student authorization, a Minister's permit, or who were refugee claimants at the time of Census (and family members living with them). Non-permanent residents accounted for 0.7, 0.4 and 0.5 percent of the samples in 1991, 1996 and 2001, respectively, and are included in the immigrant counts for those years.

We group workers in both the Canadian and U.S. labor markets into five education categories: (1) high school dropouts; (2) high school graduates; (3) workers who have some college; (4) college graduates; and (5) workers with post-graduate education.[15] We group workers into a particular years-of-experience cohort by using potential years of experience, roughly defined by Age – Years of Education – 6. Workers are aggregated into five-year experience groupings (i.e., 1 to 5 years of experience, 6 to 10 years, and so on) to incorporate the notion that workers in adjacent experience cells are more likely to affect each other's labor market opportunities than workers in cells that are further apart. The analysis is restricted to persons who have between 1 and 40 years of experience.

Our classification system implies that there are 40 skill-based population groups at each point in time (i.e., 5 education groups × 8 experience groups). Note that each of these skill-based national labor markets is observed a number of times (6 cross-sections in Canada and 5 cross-sections in the United States). There are, therefore, a total of 240 cells in our analysis of the national-level Canadian data and 200 cells in our analysis of the U.S. data.

Remarkably, even at the level of the national labor market, the sampling error in the immigrant share attenuates the wage impact of immigration. We begin our discussion of the evidence with the Canadian data because we have access to extremely large samples of the Canadian census. Table 1 summarizes the distribution of the immigrant share variable across the 240 cells in the aggregate Canadian data. The first column of the table shows key characteristics of the distribution calculated using the large file resident at Statistics Canada. These data indicate

[15] In Canada, the term "college education" typically refers to education at 2-year post-secondary institutions, while in the US it refers to 4-year university education. Throughout the text, we use the term "college" to refer to a university-level education. The Data Appendix provides a detailed discussion of the classification of the five education groups.

that 19.1 percent of the male workforce is foreign-born in the period under study, and that the variance of the immigrant share is 0.0050.[16]

The remaining columns of the top panel show what happens to this distribution as we examine progressively smaller samples of the Canadian workforce. In particular, we calculate the distribution of the immigrant share when we use data sets that comprise a 5/100 random sample of the Canadian population, a 1/100 random sample, a 1/1000 random sample, and a 1/10000 random sample. For each of these sampling rates, we drew 500 random samples from the large Statistics Canada files, and the statistics reported in Table 1 are averaged across the 500 replications. One of the replications reported in the table is of particular interest because it is the sampling rate used by Statistics Canada when they prepare the publicly available PUMF (roughly a 1 to 3 percent sample throughout the period). We drew 500 replications using the PUMF sampling rate and also report the resulting statistics.

Before proceeding to a discussion of the shifts that occur in the distribution of the immigration share variable as we draw progressively smaller samples, it is worth noting that seemingly large sampling rates (e.g., those available in the PUMF) generate a relatively small sample size for the average cell even at the level of the *national* Canadian labor market. Put differently, because the Canadian population is relatively small (31.0 million in 2001), national-level studies that calculate the immigrant share using the publicly available data may introduce substantial sampling error into the analysis. For example, the large Census files maintained at Statistics Canada yield a per-cell sample size of 30,416 observations. The PUMF replications, in

[16] The regressions presented below are weighted by the number of native workers used to calculate the mean log weekly wage of a particular skill cell. This weighting helps to adjust for differences in precision in estimating the dependent variable. To maintain consistency across all calculations, we use this weight throughout the analysis (with only one exception: to give a better sense of the distribution of cells, the percentiles of the immigrant share variable reported in Tables 1 and 3 are not weighted). We also normalized the sum of weights to equal 1 in

contrast, give a per-cell sample size of 3,247 observations. The number of observations per cell declines further to 1,400 in the 1/100 replication, to 140 in the 1/1000 replication, and to 14 in the 1/10000 replication.

Not surprisingly, Table 1 shows that the mean of the immigrant share variable is estimated precisely regardless of the sampling rate used. It is notable that the variance of the immigrant share variable increases only slightly as the average cell size declines, from 0.0050 in the large files resident at Statistics Canada to 0.0051 in the 1/100 replications and to 0.0064 in the 1/1000 replications. It is tempting to conclude that because the increase in the variance of the immigrant share variable does not seem to be very large, the problem of sampling error in estimating the wage impact of immigration may be numerically trivial. We will show below, however, that even the barely perceptible increase in the variance reported in Table 1 can lead to very large numerical changes in the estimated wage impact of immigration.

The other statistics reported in Table 1 illustrate the shifting tails of the distribution of the immigrant share as we draw smaller samples. In particular, an increasing number of cells report either very low or very high immigrant shares. In the Statistics Canada files, for example, the 10^{th} percentile cell has an immigrant share of 12.3 percent. In the 1/1000 replications, the 10^{th} percentile cell has an immigrant share of 11.2 percent, so that more cells now have few, if any, immigrants. Similarly, at the upper end of the distribution, the 90^{th} percentile cell in the Statistics Canada files has an immigrant share of 36.6 percent. In the 1/1000 replication, however, the 90^{th} percentile cell has an immigrant share of 38.8 percent, so that the cells at the upper end of the distribution are now much more "immigrant-intensive."

---

each cross-section to prevent the more recent cross-sections from contributing more to the estimation simply because each country's population increased over time.

The data for the U.S. labor market tell the same story. We use the 5/100 file to draw 500 random samples for each sampling rate: 1/100, 1/1000, and 1/10000. Even though the size of the U.S. population is almost 10 times larger than that of Canada, it is not difficult to obtain samples where the cell size falls sufficiently to raise concerns about the impact of attenuation bias—even in studies of national labor markets. The 5/100 files in the United States, for instance, lead to 47,564 observations per cell. The per-cell number of observations falls to 11,746 in the 1/100 replication, to 1,175 in the 1/1000 replication, and to 117 in the 1/10000 replication.

In the United States, as in Canada, the mean of the immigrant share distribution remains constant and the variance increases only slightly as we consider smaller sampling rates. There is also a slight fattening of the tails so that more cells contain relatively few or relatively many immigrants.

Let w_{sxt} denote the mean log weekly wage of native-born men who have education s, experience x, and are observed at time t. We stack these data across skill groups and calendar years and estimate the following regression model separately for Canada and the United States:

$$w_{sxt} = \beta p_{sxt} + \sum_i \alpha_{s(i)} S_i + \sum_j \alpha_{x(j)} X_j + \sum_l \alpha_{t(l)} T_l + \sum_i \sum_j \alpha_{sx(ij)} S_i X_j + \sum_i \sum_l \alpha_{st(il)} S_i T_l + \sum_j \sum_l \alpha_{xt(jl)} X_j T_l + \varepsilon_{sxt} \quad (9)$$

where S is a vector of fixed effects indicating the group's educational attainment; X is a vector of fixed effects indicating the group's work experience; and T is a vector of fixed effects indicating the time period. The linear fixed effects in equation (9) control for differences in labor market outcomes across schooling groups, experience groups, and over time. The interactions ($S \times T$)

and ($X \times T$) control for the possibility that the impact of education and experience changed over time, and the interaction ($S \times X$) controls for the fact that the experience profile for a particular labor market outcome may differ across education groups. Note that the regression specification in (9) implies that the labor market impact of immigration is identified using time-variation within education-experience cells. The standard errors are clustered by education-experience cells to adjust for possible serial correlation. The regressions weight the observations by the sample size used to calculate the log weekly wage. We also normalized the sum of weights to equal one in each cross-section.

Table 2 reports our estimates of the coefficient β in the Canadian labor market. Column 1 presents the basic estimates obtained from the very large files maintained by Statistics Canada. The coefficient is -0.507, with a standard error of 0.202.[17] We also estimated the auxiliary regression of the immigrant share on all the other regressors in equation (9). The *R*-squared of this auxiliary regression (reported in row 4) was 0.967, suggesting that the attenuation bias caused by sampling error could easily play an important role in the calculation of the wage impact of immigration even for relatively large samples.

We then estimated the regression model in each of the 500 randomly drawn samples for each sampling rate, and averaged the coefficient $\hat{\beta}$ across the 500 replications. The various columns of Table 2 document the impact of sampling error as we estimate the same regression model on progressively smaller samples.

---

[17] It is easier to interpret this coefficient by converting it to a wage elasticity that gives the percent change in wages associated with a one-percent immigration-induced change in labor supply. Borjas (2003, pp. 1348-1349) shows that this elasticity equals $\beta(1-p)^2$. Since the average immigrant share is around 0.2 for Canada, the coefficients reported in Table 2a can be interpreted as wage elasticities by multiplying the coefficient by approximately 0.6.

Consider initially the sampling rate that leads to the largest cell size: a random sample of 5/100 (proportionately equivalent to the largest samples publicly available in the United States). As Table 2 shows, the estimated wage impact of immigration falls by 7.7 percent; the coefficient now equals –0.468 and has an average standard error of 0.196.[18] Even when the immigrant share is calculated using an average cell size of 7,001 persons, therefore, sampling error has a numerically noticeable effect on the estimated wage impact of immigration.

The attenuation becomes more pronounced as we move to progressively smaller samples. Consider, in particular, the results from the 500 replications that use the PUMF sampling rate. Recall that this is the largest sampling rate that is publicly available in Canada. The average estimated coefficient drops to –0.403 (or a 20.5 percent drop from the estimate in the far larger Statistics Canada files). The typical researcher using the largest publicly available random sample of Canadian workers would inevitably conclude that immigration had a much smaller numerical impact on wages.[19] In fact, we can drive the estimate of β to zero by simply taking smaller sampling rates. The 1/1000 replication uses 140 observations per cell to calculate the immigrant share variable. The average coefficient is -0.076, with an average standard error of 0.191. The 1/10000 replication has 14 observations per cell and the average coefficient is -0.011, with an average standard error of 0.200.

---

[18] Note that the average standard error (across the 500 replications) is always larger than the standard deviation of the estimated coefficient across the 500 replications. We suspect that part of this difference arises because of the conservative approach that STATA uses when it computes clustered standard errors.

[19] This is not idle speculation. Bohn and Sanders (2005) replicate the national-level Borjas framework on the publicly available Canadian data and conclude that immigration has little impact on the Canadian wage structure. If we estimate the model on *the* replication that is, in fact, publicly available, the estimated coefficient is -0.210, with a standard error of 0.191. It is worth noting that, in addition to the increased sampling error, there are other notable differences between the Statistics Canada file and the publicly available PUMF. In particular, the detailed information that is provided for many of the key variables (e.g., years of schooling and labor force activity) in the Statistics Canada file is not available in the PUMF file because the values for some variables are reported in terms of intervals.

It is easy to show that the substantial drop in the estimated wage impact of immigration as we move to progressively smaller random samples can be attributed to sampling error. Because we have access to the "true" immigrant shares in Canada (i.e., the immigrant shares calculated from the large Statistics Canada files), we can correct for sampling error by simply running a regression that replaces the error-ridden measure of the immigrant share with the true immigrant share in each of our replications. The distribution of the coefficient from this regression, β^*, is reported in rows 6-8 of Table 2.

In every single case, regardless of how small the sampling rate is, we come very close to estimating the "true" coefficient—although there is a great deal of variance in the estimated wage impact across the replications. In particular, the coefficient estimated in the Statistics Canada file is -0.507. If we used the correct immigrant share in the 1/100 replications the estimated coefficient β^* is -0.499, and the standard deviation of this coefficient across the 500 replications is 0.126. Similarly, if we used the correct immigrant share in the 1/1000 replication, the estimated coefficient is -0.466, and the standard deviation of this coefficient is 0.405. Even in the 1/10000 replication, with only 14 observations per cell, the use of the "true" immigrant share leads to a coefficient that is much closer to the true wage impact (although it is very imprecisely estimated): the coefficient is -0.384, with a standard deviation of 1.353. In sum, Table 2 provides compelling evidence that sampling error in the measure of the immigrant share can greatly attenuate the estimated wage impact of immigration.[20]

Of course, the typical analyst will not have access to the "true" immigrant share in the Statistics Canada file so that this method does not provide a practical way for calculating

[20] Note that the estimates constructed using the "true" immigrant share fall with sample size. As the sample size gets smaller, the classification errors discussed in section 2 may become increasingly more important if the

consistent regression coefficients. It is crucial, therefore, to consider alternative methods of correcting for attenuation bias. Equation (7) provides a simple solution to the problem as long as the measurement error is attributable solely to sampling error and no other variables are measured with error.[21] In particular, equation (7) allows us to conduct a back-of-the-envelope calculation of what the coefficient β would have been in the absence of sampling error. This exercise requires information on the immigrant share in the population, the observed variance of the immigrant share, the R^2 from the auxiliary regression, and average cell size. We calculated the corrected coefficient for each of the 500 replications at each sampling rate. Row 9 of Table 2 reports the average corrected coefficient and row 10 reports the standard deviation across replications.[22]

The corrected coefficients reported in Table 2 reveal that *even* the coefficients estimated using the large files resident at Statistics Canada are not immune to sampling error. Although the bias is not large, using either of the correction methods described above suggests that the "true" wage impact of immigration in Canada is -0.52, implying an attenuation bias of 2.5 percent even with a cell size of over 30,000 persons.

---

noise to signal ratio becomes larger. This may attenuate the estimated parameters (see Freeman, 1984; and Paggiaro and Torelli, 2004).

[21] Some of the replications combine samples collected at different sampling rates. The sampling rate is set at 0.20 for the corrections in the Statistics Canada file; 0.025 for the corrections in the PUMF replication; and 0.05 for the corrections in the 5/100 file for the United States.

[22] As an alternative to the back-of-the-envelope method for calculating the coefficient β, we can instead estimate the mean of the sampling error defined in equation (4) by:

$$E\left(\frac{\pi_k(1-\pi_k)}{n_k}\right)=\frac{\sum_k \lambda_k p_k(1-p_k)/n_k}{\sum_k \lambda_k},$$

where the weight λ_k gives the number of native workers in cell *k.* This procedure involves the calculation of this expectation for each of the 500 replications in our simulation. The results from this alternative procedure, not reported here, are similar to those reported in rows 9 and 10 of Table 2 produced by the back-of-the-envelope calculation.

The back-of-the-envelope correction generates adjusted coefficients that typically approximate this "true" effect as long as the mean cell size is large, but is much less useful when the mean cell size declines. A useful "rule of thumb" seems to be that one needs at least 1,000 observations per cell in order to predict the true coefficient with some degree of accuracy. In the 5/100 replications, for example, the adjusted coefficient is around -0.53. At the PUMF sampling rate, the inconsistent coefficient $\hat{\beta}$ is -0.403. The adjusted coefficient is -0.52 if we use the back-of-the-envelope approach in equation (7). The adjustment goes further off the mark if we move to the 1/100 replications. The estimate is -0.64, with a very large standard deviation. Finally, if the cell size gets sufficiently small, as in the 1/1000 replication, the back-of-the-envelope correction breaks down. At this sampling rate, the predicted amount of sampling error often exceeds the adjusted variance of the observed immigrant share, leading to very unstable corrections.

As noted above, the back-of-the-envelope correction, although easily applied in practice, may be misleading because the usual OLS estimator of β that is used in the correction is itself biased. As a result, equation (8) may not provide a good estimate of the size of the attenuation bias. It is therefore important to compare this method to alternative, relatively more complex, methods that provide consistent estimates of the parameter of interest β. The first such consistent estimator, as discussed by Greene (1993, p. 282), can be obtained if $Var(u_k)$ is known, or can be approximated. As we show in the Mathematical Appendix, an alternative estimator of β in the current context is given by:

$$\tilde{\beta} = \frac{\widehat{Cov}(w_k, r_k)}{\widehat{Var}(r_k) - \left(\frac{K-H}{K}\right)\widehat{Var}(u_k)}, \qquad (10)$$

where r_k are the residuals from a regression of p_k on Z_k in equation (1), where Z_k includes a constant and all the fixed effects and interactions defined in equation (9), K is the number of labor markets, and H is the dimension of the vector Z_k. In short, the estimation strategy involves a first stage regression of p_k on Z_k, which provides the residual r_k, and then computing the moments that appear in the numerator and denominator of equation (10). Following the strategy introduced in the previous section, $\widehat{Var}(u_k) \approx (1-\tau)\frac{\bar{p}(1-\bar{p})}{\bar{n}}$ and the remaining two terms in equation (10) can be calculated directly from the data. While consistent, the estimator $\tilde{\beta}$ is subject to a finite sample bias because of the nonlinear transformations applied to the unbiased estimators of the numerator and the denominator, and a correction for this bias is possible by using a Monte Carlo procedure for Bootstrap Bias Estimation (BBE) (Horowitz, 2001). The BBE procedure involves the following steps:

Step 1: Use the data from the estimation sample and the approximation for $Var(u_k)$ to compute $\tilde{\beta}$.

Step 2: Generate a bootstrap sample of exactly the same size as the original microdata by sampling randomly with replacement. We then use the new estimation sample to compute $\tilde{\beta}^*$. We draw 500 such bootstrap samples for each of the replications.

Step 3: Compute $E[\tilde{\beta}^*]$ by averaging the results of the 500 repetitions in step 2.[23] We then define the bias as $B^* = E[\tilde{\beta}^*] - \tilde{\beta}$.

Step 4: The bias-corrected estimator of β is then given by $\tilde{\beta} - B^*$.

Row 11 of Table 2 presents the results of the BBE method. The bias-corrected estimates are very similar to those obtained using the back-of-the-envelope correction for large samples. For the 5/100 sampling rate, for instance, the BBE estimate of β is -0.522, as compared to -0.531 with the back-of-the-envelope correction. It is worth noting that the BBE method also breaks down when the sample size becomes very small.

Finally, it is of interest to compare these bias-corrected, consistent estimates to those obtained from an alternative method based on instrumental variables. The IV approach for correction of attenuation bias, first proposed by Griliches and Mason (1972), requires that we observe two measures of the variable subject to error. The two measures have the property that they are correlated with each other, but have uncorrelated measurement errors. The second measure is then used as an instrument for the first to correct for the attenuation bias.

We employ the unbiased split sample instrumental variable (USSIV) method to correct for attenuation bias (Angrist and Krueger, 1995). In our context, this method boils down to splitting each sample randomly into half samples and using observed immigrant shares from the second-half sample as instruments in the first-half sample. More formally, for a given replication we first split the sample randomly into two parts. For labor market *k*, let p_k^1 and p_k^2 be the observed immigrant shares in the first- and second-half samples. Both p_k^1 and p_k^2 are measures of the true immigrant share such that $p_k^1 = \pi_k + u_k^1$ and $p_k^2 = \pi_k + u_k^2$. For a given labor market *k*, p_k^1 and p_k^2 are correlated, but the measurement errors u_k^1 and u_k^2 are uncorrelated because the half samples are drawn randomly. We then use the data from the first half sample to estimate:

---

[23] We also conducted some of the simulations using 2000 repetitions at this stage rather than 500 with the Canadian data and the results were similar to the ones reported here.

$$w^1_{sxt} = \beta\, p^1_{sxt} + \sum_i \alpha_{s(i)} S_i + \sum_j \alpha_{x(j)} X_j + \sum_t \alpha_{t(l)} T_l + \sum_i \sum_j \alpha_{sx(ij)} S_i X_j + \sum_i \sum_l \alpha_{st(il)} S_i T_l + \sum_j \sum_l \alpha_{xt(jl)} X_j T_l + \varepsilon_{sxt} \quad (11)$$

and instrument p^1_{sxt} with p^2_{sxt}.

For a given sampling rate, we estimated equation (11) for each of the 500 replications. We also applied this method in the Statistics Canada file by creating 500 half sample pairs from the Statistics Canada file using a random number generator with different seeds and then estimating the USSIV corrected coefficients for each case.[24] The estimated USSIV regression coefficients are reported in row 12 of Table 2 (and the standard deviation is reported in row 13). For larger sampling rates, the USSIV estimates are very similar to those estimated using the other correction methods. Consider, for instance, the results obtained in the PUMF replication. The OLS coefficient estimated using the mismeasured immigrant share variable is -0.403; the back-of-the-envelope correction yields a bias-corrected estimate of -0.524; and the USSIV method yields an estimate of -0.510.[25] Note, however, that the USSIV method also breaks down as the cell size becomes smaller. In the 1/1000 replication, for example, the mean USSIV coefficient changes sign and becomes 0.482.[26] As a general rule, the various (and very different)

---

[24] In the U.S. context, the analogous procedure is to create 500 half-sample pairs from the 5/100 data using a random number generator with different seeds and then estimate the USSIV corrected coefficients for each case.

[25] It is also possible to use instruments based on the economics of the model, rather than the purely statistical approach in USSIV, to correct for sampling error bias. We will discuss below the problems introduced by sampling error when one uses the preferred instrument in the literature, a lagged measure of the immigrant share in labor market *k*.

[26] The average coefficient across the 500 replications is generally similar to the median for sufficiently large sampling rates. In the Canadian data, for example, the mean and median estimates for the 1/100 sampling rate are -0.525 and -0.510 respectively. The mean and median estimates, however, are 0.482 and -0.302 for the 1/1000 sampling rate, and -0.486 and -0.134 for the 1/10000 sampling rate.

methods of correction for attenuation bias tend to work only when the average cell in the Canadian national labor market has at least 1,000 observations.

Table 3 replicates the analysis using the data available for the U.S. labor market. Note that our largest sample is the publicly available IPUMS of the decennial Census—which represents a 1% sampling rate in 1960, a 3% sampling rate in 1970, and a 5% sampling rate from 1980 through 2000. The estimate of the wage impact of immigration at the national level in this large sample is quite similar to that found with the Statistics Canada data: the estimated coefficient is -0.489, with a standard error of 0.223. Note, however, that because of the much larger U.S. population, the mean cell size is far larger (47,514 observations) than the mean cell size in the Statistics Canada file (30,416 observations). Note also that applying any method of correction to the coefficient estimated in this very large U.S. sample only slightly increases the magnitude of the estimated wage impact of immigration to around -0.5.

As with Canada, we estimated the model using 500 replications for each smaller sampling rate. The 1/100 replications have 11,746 observations per cell. As a result, the estimated coefficient $\hat{\beta}$ declines only slightly. The cell size in the 1/1000 replications, however, is much smaller (1,175 observations per cell), and the estimated coefficient falls to -0.347, with an average standard error of 0.247. In other words, the bias attributable to sampling error reduces the coefficient by almost 30 percent. Studies that use this sampling rate—even if they focus on national labor market trends and have over 1,000 observations per cell—will falsely conclude that the wage impact of immigration is numerically weak and statistically insignificant. Table 3 shows that we can drive the estimated wage impact of immigration to zero by simply taking an even smaller sampling rate. The 1/10000 replication, where the average cell size used to

calculate the immigrant share variable has 117.4 workers, has an average coefficient of -0.082, with an average standard error of 0.279.

The hypothesis that sampling error generates exponentially smaller immigration effects as we use smaller samples is confirmed by the regressions that use the "true" immigrant share (i.e., the immigrant share calculated from the 5/100 files). The coefficient β^* estimated in these regressions is reported in row 6 of the table. The estimated coefficients using the more precise measure of the immigrant share tend to almost exactly duplicate the estimated wage impact obtained from the 5/100 file. Even in the 1/10000 replication, where the wage impact of immigration estimated with the error-ridden immigrant share variable is essentially zero, the use of the immigrant share from the 5/100 file raises the coefficient to -0.498, almost exactly what we obtained in the "population" regression (although it is imprecisely estimated).

The remaining rows of the table show what happens to the estimated wage impact of immigration when we use the three alternative correction methods to adjust the inconsistent estimate for sampling error. It turns out that only the USSIV method leads to a sensible estimate in the 1/1000 replication and none of the correction methods lead to sensible estimates in the 1/10000 replication. As with the Canadian data, all of the correction methods break down when the average cell size in the U.S. national labor market falls below 1,000 observations.

B. Local Labor Markets

Up to this point, we have considered national labor markets defined in terms of skills (education and experience). We now adopt the convention used in much of the literature and consider labor markets (within skill groups) defined by the geographic boundaries of metropolitan areas. There are approximately 27 identifiable metropolitan areas in each Canadian

census beginning in 1981, and over 250 identifiable metropolitan areas in the U.S. Census beginning in 1980.[27] Workers who do not live in one of the identifiable metropolitan areas are excluded from the analysis. Because labor markets are now defined in terms of metropolitan area, education, experience, and time, the number of cells increases dramatically. There are 5,360 cells in Canada and 31,472 cells in the United States.[28] It immediately follows that the number of observations per cell declines substantially once we move the unit of analysis to this level of geography.[29]

Table 4 reports the distribution of the immigrant share variable estimated at the metropolitan area level for both Canada and the United States. In Canada, the per-cell number of observations is 660 even when we use the large confidential files maintained by Statistics Canada. If we use the PUMF sampling rate, the average cell contains only 84 observations. By the time we use the 1/100 sampling rate, we only have 34 observations per cell. In the United States, the 5/100 Public Use Samples yields only 174 observations per cell, and this number drops to just 36 observations if we use a 1/100 sampling rate. Because even the 1/100 sample in

---

[27] The census file maintained at Statistics Canada identifies 26 metropolitan areas in the 1981 Census and 27 metropolitan areas in each census since 1986. The publicly available PUMF identifies far fewer metropolitan areas; in 2001, for example, only 19 metropolitan areas are identified in the public file. The IPUMS file of the U.S. Census identifies 255 metropolitan areas in 1980, 249 metropolitan areas in 1990, and 283 metropolitan areas in 2000. The definition of the metropolitan areas in both the Canadian and U.S. censuses is substantially different prior to 1980, so our analysis of wage differences across local labor markets is restricted to the census data that begins in 1980/1981.

[28] The number of cells in our analysis of the 5/100 file in the United States is slightly smaller than the theoretically possible number of cells (31,480) because there are a few empty cells—that is, there are labor markets where we could not detect any native working men. These labor markets are not included in the regressions and create an additional source of error in estimates of the wage impact of immigration. This error will obviously be more important for smaller sampling rates.

[29] Although the per-cell size is much smaller in the spatial correlation analysis than in the national labor market analysis, we show below that the variance of the observed immigrant share across labor markets is much higher. This large variance suggests that the estimated wage impact of immigration at the local level—for a given cell size—would be less attenuated by sampling error than the comparable estimate at the national level.

Canada and the 1/100 sample in the United States have very few observations per cell, we limit our analysis of spatial correlations to sampling rates that are at least as large as these.

As in the previous section, there is little difference in the mean immigrant share across the various sampling rates, and only a slight increase in the variance of the immigrant share variable as we use smaller samples. However, the small increase in the variance masks a substantial increase in the number of cells that have no immigrants as we use progressively smaller samples in either country.

We use the following regression specification to estimate the wage impact of immigration in local labor markets. Let w_{hrt} denote the mean log weekly wage of native men who have skills h (i.e., a particular education-experience combination), work in metropolitan area r, and are observed at time t. For each country, we stack these data across skill groups, geographic areas, and Census cross-sections and estimate the model:

$$w_{hrt} = \beta p_{hrt} + \sum_i \alpha_{h(i)} H_i + \sum_j \alpha_{r(j)} R_j + \sum_l \alpha_{t(l)} T_l + \sum_i \sum_j \alpha_{hr(ij)} H_i R_j + \sum_i \sum_l \alpha_{ht(il)} H_i T_l + \sum_j \sum_l \alpha_{rt(jl)} R_j T_l + \phi_{hrt} \quad (12)$$

where H is a vector of fixed effects indicating the group's skill level; R is a vector of fixed effects indicating the metropolitan area of residence; and T is a vector of fixed effects indicating the time period of the observation. The standard errors are clustered by skill-region cells to adjust for the possible serial correlation that may exist within cells.

Table 5 reports the coefficients estimated for the various specifications. It is well known that because labor or capital flows across metropolitan areas arbitrage geographic wage

differences, the labor market impact of immigration estimated at the metropolitan area level will typically be smaller than that estimated at the national level—even in the absence of attenuation bias. Therefore, it is not surprising that the coefficient $\hat{\beta}$ reported in Table 5 is substantially smaller than that found in the national-level analysis even when we use the largest samples available. In Canada, for example, the estimated effect using the Statistics Canada file is -0.053, with a standard error of 0.037. In the United States, the estimated effect is remarkably similar; the coefficient is -0.050, with a standard error of 0.023.

Before we turn to the various replications, it is worth noting that because the sample size used to calculate the immigrant share variable is relatively small even using these large samples, the estimated wage effect of approximately -0.05 in either country may have already been greatly attenuated by sampling error.[30] The corrected coefficients reported in the table confirm our suspicions. Row 8 of Table 5 shows that the simplest back-of-the-envelope correction *more than doubles* the estimated wage impact to -0.112 in Canada, so that the bias in the spatial correlation using the large Statistics Canada file is around 53 percent. Similarly, the back-of-the-envelope correction in the United States *more than triples* the estimated wage impact to -0.170 in the United States, implying a bias of around 70 percent. The use of BBE or USSIV leads to roughly

[30] Card's (1991) influential study of the Mariel flow is not susceptible to the type of sampling error documented in this paper. Card compares labor market conditions in Miami and a set of other cities before and after the Mariel flow of immigrants in 1980. He finds little change in Miami's labor market conditions (relative to the comparison cities) during the period. The interpretation of Card's evidence, however, is very unclear. Angrist and Krueger (1999) replicated Card's study by examining conditions in Miami and the same comparison cities in 1994. The 1994 period is notable because conditions in Cuba were ripe for the onset of a new wave of refugees, and thousands of Cubans began the hazardous journey. The Clinton administration, however, rerouted all the refugees towards the American military base in Guantanamo Bay, so few of the potential migrants arrived in the U.S. mainland by 1995. Remarkably, Angrist and Krueger's replication finds that a phantom immigrant influx ("The Mariel Boatlift That Didn't Happen") had a *significant and adverse* impact on labor market conditions in Miami. At least in 1994, there were confounding factors in Card's difference-in-differences methodology that are not well understood and drive the results.

similar conclusions: The corrected estimate of the coefficient β in Canada or the United States is almost double the size of the OLS regression coefficient.

Not surprisingly, the bias in the estimated wage impact of immigration becomes substantially worse when we consider smaller samples. In the PUMF sampling rate, the average estimated wage impact of immigration at the metropolitan area level is only -0.022, with an average standard error of 0.039. The publicly available data, therefore, leads to a completely different substantive conclusion (i.e., no statistically significant wage impact of immigration at the local level) than the larger Statistics Canada file. As row 5 of the table shows, however, we can replicate the impact implied by the Statistics Canada data (-0.053) in the PUMF replications if we had used the immigrant share that can be calculated in the large Statistics Canada sample. Because the average cell size becomes very small, the precision of our corrected coefficients declines dramatically as we use smaller sampling rates.

The analysis of wage differences across local labor markets in the United States leads to very similar results. As noted above, we only consider one sampling rate because even at the 1/100 level there are only 36 observations per cell. The average wage impact of immigration estimated in the 1/100 replications is less than half the size of that estimated using the larger 5/100 files; the average coefficient is -0.022, and the average standard error is 0.027. As in Canada, the use of the 1/100 sampling rate would lead researchers to conclude that the wage impact of immigration at the local level is numerically and statistically zero, when in fact a different conclusion would have been reached if the analyst had used a much larger sample.

Because of the influence of the spatial correlation approach in the literature, it is of interest to discuss the implications of the results presented in this section for the existing evidence. The estimated effects reported in Table 5 define the labor market along the lines of

both skills and geography, using 40 skill groups (5 education groups and 8 experience groups) and every single metropolitan area that can be identified by the Canadian or U.S. Census. There are 283 identifiable metropolitan areas in the U.S. Census, so that there are potentially 11,320 cells in each cross-section. Inevitably, the large number of distinct labor markets leads to a small average cell size. In an important sense, the research design used in Table 5 introduces the possibility that sampling error could play an important role in estimates of the spatial correlation.

Many studies in the spatial correlation literature, however, avoid the problem of small cell size by altering the research design in one of two distinct ways. The first, used mainly in the earlier (though influential) studies in the literature, is simply to calculate the immigrant share in the metropolitan area—and completely disregard the skill composition of the local immigrant population in the calculation. For example, Altonji and Card (1991, p. 217) define the immigrant share as "the fraction of foreign-born residents" in each metropolitan area. Similarly, Schoeni (1997, p. 7) defines the immigrant share as "the share of the entire working aged population in that area who are immigrants; the sample sizes in the Census are too small to calculate sub-group specific shares in most geographic areas."[31] These studies then proceed to relate the wage of specific groups (such as low-skill natives, or low-skill blacks) to the aggregate immigrant share in the locality.

This research strategy, however, introduces a different problem into the analysis. Because the wage of, say, low-skill workers is being correlated with the immigration-induced *total* supply shift in the locality, it is unclear what structural parameter this correlation is supposed to

[31] Altonji and Card (1991, p. 217) note that "since our sample sizes for 1970 are too small to provide reliable estimates of the fraction of immigrants in many of the smaller cities, we have relied instead on the published population data." The use of the published Census data implies that sampling error is unlikely to be an issue in the Altonji-Card analysis. Schoeni (1997, p. 7), however, estimates the immigrant share "internally from the census analysis files."

measure. In some metropolitan areas, for example, the immigrant influx is disproportionately low-skill, and one might expect to find a negative correlation between the wage of the pre-existing low-skill workforce and the immigrant share. In other metropolitan areas, however, the immigrant influx may be disproportionately high-skill, and there could potentially be a positive correlation between the wage of low-skill workers and the immigrant share because of production complementarities across skill groups. The sign of the net correlation estimated in the data, therefore, is unpredictable and depends on the settlement patterns of low-skill natives and low-skill immigrants. Although these studies avoid the attenuation bias introduced by sampling error, they do not estimate any parameter of interest.

Another common method for avoiding sampling error is to limit the analysis to the largest metropolitan areas in the country. Card (2001), for example, uses only the 175 cities with the largest number of native-born adults in the population, while Butcher and Card (1991, p. 292) use only 24 cities, which include "the 10 most immigrant-intensive cities." This approach would be sensible if immigrants settled randomly across metropolitan areas in the United States. The assumption of random settlement, however, is clearly false. In 2000, for example, 38.4 percent of immigrants lived in four metropolitan areas (New York, Los Angeles, Chicago, and San Francisco), but only 12.2 percent of natives lived in the four metropolitan areas with the largest native-born populations (New York, Chicago, Los Angeles, and Philadelphia). Even abstracting from the econometric problems potentially introduced by non-random sample selection, the spatial correlations obtained from a selected sample of cities may have limited applicability. In particular, they may provide internally valid estimates of the wage impact of immigration for those cities, but may provide little information about the wage impact of immigration across the entire labor market.

Moreover, it is easy to show that although restricting the analysis to the largest cities does indeed increase average cell size, these larger samples come at a cost. The analyst is effectively introducing a specific type of selection bias into the analysis.[32] Although we are unaware of any study that examines the magnitude of the bias generated by the non-random sampling of local labor markets, it is easy to show that the resulting bias is numerically important. As Table 5 shows, the coefficient in the United States is -0.050 (0.023) when the regression uses the full set of 283 metropolitan areas available in the 5/100 Census microdata. Remarkably, this coefficient falls to -0.013 (0.034) if we estimate the regression using only the data from the largest 50 metropolitan areas; to -0.041 (0.028) if we use the largest 100 metropolitan areas; to -0.042 (0.025) if we use the largest 150 metropolitan areas; and to -0.045 (0.024) if we use the largest 200 metropolitan areas.[33] In sum, restricting the analysis to the largest metropolitan areas reduces the size of the coefficient by 10 to 20 percent even when the regression uses half or two-thirds of the available local labor markets. This way of avoiding sampling error introduces a different type of bias—a selection bias that also leads to a numerically meaningful attenuation of the wage impact of immigration.

C. Using Lagged Immigration as an Instrument

Income-maximizing immigrants may cluster in particular (geographic or skill-based) labor markets because those are the markets that offer relatively high returns to the mobility costs incurred by the migrants. The immigrant share coefficient from an OLS wage regression

---

[32] For instance, suppose that immigration has an adverse wage impact on competing workers. All other things equal, the sample selection rule used in these studies omits from the analysis those metropolitan areas with the highest wages, so that the regression introduces a type of selection bias based on selection in the dependent variable.

[33] The labor markets are ranked by the size of the native-born male workforce in 2000.

would then be positively biased. Some studies use instrumental variables to account for this potential endogeneity problem (e.g., Altonji and Card, 1991; Schoeni, 1997; Card, 2001). The typical instrument is some lagged measure of the immigrant share, on the presumption that the continuing influx of immigrants into particular markets is based mostly on the magnetic attraction of network effects rather than on any income-maximizing behavior.[34] In theory, these IV regressions provide an alternative method for correcting for sampling error bias because the sampling error in the current and lagged values of the immigrant share is uncorrelated in independent samples.

Of course, it is far from clear that the lagged immigrant share is a legitimate instrument: What factors attracted large numbers of particular immigrants to particular markets in the first place? If the earlier immigrant arrivals selected those markets *because* they offered relatively better job opportunities, any serial correlation in these opportunities violates the orthogonality conditions required in a valid instrument. Even abstracting from this conceptual question, it turns out that the sampling error in the immigrant share variable creates serious statistical problems for *this* particular instrument, leading both to weak instruments and to the violation of a key assumption in the classical measurement error model.

Consider the generic first-stage regression:

$$p_{tk} = \delta\, p_{t-1,k} + \sum_h \gamma_h z_{tkh} + \varepsilon_{tk}\,, \tag{13}$$

[34] Although the IV methodology has been used exclusively in studies conducted at the metropolitan area level, a similar type of argument suggests that the lagged immigrant share could serve as an instrument in national-level studies as well. Immigration policy in both Canada and the United States, for example, gives entry preference to family members of persons already residing in the receiving country. If skill levels are correlated within families (e.g., spouses and siblings may have roughly the same age and education level as the visa sponsor), an immigrant influx in a particular skill group at time *t* would likely generate more immigrants with similar skills in the future.

where p_{tk} is the observed immigrant share for cell k in the current period and $p_{t-1,k}$ is the lagged share. The observed immigrant shares are defined by: $p_{tk} = \pi_{tk} + u_{tk}$ and $p_{t-1,k} = \pi_{t-1,k} + u_{t-1,k}$, where the sampling errors have mean zero, are uncorrelated with the true immigrant share, and are uncorrelated over time. The vectors of fixed effects included in the first-stage regression are the same as those included in equation (9) for the national-level analysis and equation (13) for the metropolitan area analysis.[35]

Table 6 summarizes the results of our sensitivity analysis of the first-stage regression model. The qualitative nature of the evidence is very similar for both Canada and the United States. The coefficient of the lagged immigrant share in the large Statistics Canada file is 0.258, with a standard error of 0.085, implying that the *F*-statistic associated with the instrument is 9.21, very close to the threshold (an *F*-statistic above 10) required to reject the hypothesis that the lagged immigrant share is a weak instrument (in the sense defined in Stock, Wright, and Yogo, 2002). Initially, as we consider smaller sampling rates, the estimated coefficient $\hat{\delta}$ goes towards zero, and the lagged immigrant share becomes an obviously weak instrument. In the replications using a 1/100 sampling rate, for example, the coefficient is 0.054 and the standard error is 0.100. As the cell size gets smaller still, however, the coefficient $\hat{\delta}$ turns very negative and significant! Note that this sign reversal occurs in the national level regressions for both Canada and the United States, as well as in the metropolitan area regressions for Canada. In the

[35] There is a 10-year gap between the 1971 and 1981 Canadian cross-sections, but only a 5-year gap between the post-1981 censuses. To ensure that the lagged immigrant share is defined consistently, we omit all cells from the 1971 Canadian census in the regressions reported in this section. As a result, the first-stage regressions estimated in Canada only include cells beginning with the 1986 cross-section. The national level regressions for the United States include cells beginning with the 1970 census, and the metropolitan area regressions for the United States include cells beginning with the 1990 census. Finally, all the models estimated at the metropolitan area level include only those metropolitan areas that are identified in each cross-section.

metropolitan area analysis for the United States, the coefficient $\hat{\delta}$ is *already* negative even at the 5/100 sampling rate.[36] In short, the first-stage IV regression seems to completely break down when the immigrant share is calculated in relatively small samples.

It is easy to demonstrate that the IV method fails because the sampling errors on both sides of the first-stage regression equation are correlated. Table 6 reports the average estimate of two other regression coefficients: $\hat{\delta}(p_t, \pi^*_{t-1})$, which is the coefficient obtained by regressing the observed current immigrant share on the "true" lagged immigrant share (i.e., the share calculated in the largest available sample—either the Statistics Canada file or the 5/100 U.S. Census); and $\hat{\delta}(\pi^*_t, p_{t-1})$, which is the coefficient from the regression of the "true" current immigrant share on the observed lagged share. Note that the average value of $\hat{\delta}(p_t, \pi^*_{t-1})$ often replicates the positive and sizable coefficient obtained when the regression is estimated in the largest file available, confirming that sampling error in the dependent variable does not typically affect the regression coefficient. Similarly, the average of $\hat{\delta}(\pi^*_t, p_{t-1})$ is often close to zero, confirming that sampling error in the independent variable attenuates the estimated coefficient.

We find negative and significant estimates of δ only when both the current and the lagged immigrant share are measured with substantial sampling error. Although it would seem that the errors are uncorrelated because sampling error is independent across cross-sections, the first-stage regression model actually builds in a strong negative correlation in the errors between the

---

[36] Despite the fact that the lagged immigrant share enters the regression with the wrong sign, some of the regression specifications reject the hypothesis that the lagged immigrant share is a weak instrument. It is well known, however, that standard IV specification tests have no power to detect the problems associated with the type of non-classical measurement error documented in this section (Kane, Rouse, and Staiger, 1999).

two sides of the equation.[37] In particular, the fixed effect specification effectively differences the data from the mean immigrant share observed in labor market *k* during the sample period (where labor market *k* is defined by skill and/or geography). As a result, we can write the first-stage regression model in its equivalent differenced form as:

$$p_{t,k} - \overline{p}_{t,k} = \delta(p_{t-1,k} - \overline{p}_{t-1,k}) + \text{fixed effects} + \varepsilon, \tag{14}$$

where $\overline{p}_{t,k}$ is the average of the current immigrant share across the various cross-sections available for the labor market, and $\overline{p}_{t-1,k}$ is the corresponding average of the lagged immigrant share. The implications of this type of differenced structure for correlated sampling errors are obvious by considering the special case where the data consists of two cross-sections, and the fixed effect estimator is equivalent to a simple differencing of the data.[38] We can then rewrite equation (14) as:

$$p_{tk} - p_{t\text{-}1,k} = \delta\,(p_{t\text{-}1,k} - p_{t\text{-}2,k}) + \text{fixed effects} + \varepsilon. \tag{14′}$$

---

[37] In fact, the measurement errors seem to be themselves serially correlated over time. This correlation might arise if there are persistent differences in the quality of local census operations or in the kinds of informal living arrangements that make enumeration difficult. It is easy to investigate this possibility in the Canadian context. In particular, we estimated MSA-level regressions of the observed immigrant share at a given sampling rate on the size of the MSA population, fraction of the population in the MSA in each education category (5 categories), and experience category (8 categories), separately by survey year. We then calculated the measurement error as the difference between the "true" immigrant share (obtained from the Statistics Canada file) and the one predicted from the MSA-level regression. We used the 1986, 1991, 1996, and 2001 Censuses to carry out a Monte Carlo study with 500 random samples for each sampling rate. At the 5 in 100 sampling rate, the regression of the predicted 2001 measurement error on the predicted measurement errors for previous years leads to an average coefficient of 0.619 (mean standard error of 0.04) for the 1996 measurement error. The corresponding estimates for the 1991 and 1986 predicted measurement errors are 0.09 (0.04) and 0.17 (0.03), respectively.

[38] Note, however, that three cross-sections are required to actually estimate the model..

The appearance of $p_{t\text{-}1,k}$ on both sides of the equation indicates that any sampling error in the regressor gets completely transmitted—*with a negative sign*—to the dependent variable, violating one of the key assumptions of the classical measurement error model. The negative correlation between the measurement errors in the dependent and independent variables in (14′) imparts a substantial negative bias on the coefficient δ when there is sufficiently large sampling error in the observed immigrant share.

The insight that the first-stage regression can be interpreted as a first-difference regression helps explain the pattern of estimated coefficients reported in Table 6. In particular, note that in the apparent absence of sampling errors (e.g., in the national-level regressions estimated either in the Statistics Canada or 5/100 U.S. Census files), the estimated coefficient $\hat{\delta}$ is strongly positive. Errors in the right-hand-side of equation (14′) attenuate the coefficient towards zero, while errors in the left-hand-side have relatively little influence on the estimate. However, the existence of negatively correlated errors on both sides of the equation turns the estimated coefficient strongly negative. The results summarized in Table 6 clearly indicate that the lagged immigrant share is not a good instrument when the cell size is sufficiently small—even when we abstract from any conceptual issues.

IV. Summary

The parameter measuring the wage impact of immigration plays a crucial role in any discussion of the costs and benefits of immigration on a receiving country. Because of its importance, a large and influential empirical literature developed over the past 20 years. Although economic theory predicts that the relative price of labor would decline as a result of the immigration-induced supply increase (at least in the short run), many studies, particularly those

that use geographic variation in wage levels to measure the relation between wages and immigration, conclude that the wage impact of immigration is negligible.

This paper tests a new hypothesis that may account for some of the weak estimated effects in the literature: the estimated wage impact of immigration is attenuated by measurement error. In particular, the key independent variable in the analysis, the fraction of the workforce that is foreign-born, is typically calculated from a sample of workers in the labor market of interest. This calculation introduces sampling error into the key independent variable and leads to attenuation bias through the usual errors-in-variables model. Sampling error plays a disproportionately large role because of the longitudinal nature of the econometric framework commonly used to measure the wage impact of immigration. After controlling for permanent factors that determine wages in labor markets, there is little variation remaining in the immigrant share.

Our analysis used labor market data drawn from both Canada and the United States to show that: (a) the attenuation bias is quite important in the empirical context of estimating the wage impact of immigration; and (b) adjusting for the attenuation bias can easily double, triple, and sometimes even quadruple the estimated wage impact of immigration. Our evidence also indicated that the attenuation bias becomes exponentially worse as the size of the sample used to calculate the immigrant share in the typical labor market declines.

In an important sense, previous research in this literature has been conducted under the false sense of security provided by the perception that the analysis is sometimes carried out using very large samples (such as the 5 percent file of the U.S. Census). The use of these large data files would seem to suggest that the immigrant share of the workforce is measured accurately. We have shown, however, that even as large a sampling rate as a 5/100 file can easily generate

substantial sampling error in the immigrant share—and that this sampling error will almost certainly be a numerically important factor in longitudinal-type studies where the labor market is defined in terms of narrow skill groups and/or geography. Measurement error, therefore, has been an important—and previously ignored—contaminant of the empirical results reported in this literature.

The false sense of security provided by the large microdata Census samples probably extends to many other contexts in applied economics. After all, there are many empirical studies where calculated proportions form the key variable of interest in a longitudinal context. Consider, for example, regression models where the key regressor is a group-specific unemployment rate or the fraction of the workforce belonging to a particular racial or ethnic group. In view of the evidence reported in this paper, it would not be far-fetched to conjecture that the conclusions of many of those studies are also likely to be very sensitive to attenuation bias. A greater appreciation for the problems introduced by binomial-based sampling error in independent variables could lead to a reappraisal of many regression-based stylized facts.

Mathematical Appendix

A. Proof of equation (4)

For simplicity, we consider the case where the sampling rate τ approaches zero so that we can refer exclusively to the properties of the binomial distribution. The extension to the hypergeometric distribution is straightforward. The relation between the observed and true immigrant share is given by:

$$p_k = \pi_k + u_k \,. \tag{A1}$$

Note that the error in (A1) can be written as:

$$u_k = p_k - \pi_k = \frac{1}{n_k}(n_k p_k - n_k \pi_k). \tag{A2}$$

Conditional on the sample size n_k used to calculate the immigrant share and on π_k, $n_k p_k$ is a binomial random variable with parameter (n_k, π_k). The central moments of the binomial distribution imply:

$$\begin{aligned} E[u_k \mid n_k, \pi_k] &= \frac{1}{n_k}(n_k \pi_k - n_k \pi_k) = 0, \\ E[u_k^2 \mid n_k, \pi_k] &= \frac{1}{n_k^2} n_k \pi_k (1-\pi_k) = \frac{\pi_k (1-\pi_k)}{n_k}, \\ E[u_k^4 \mid n_k, \pi_k] &= \frac{1}{n_k^4}\left(3(n_k \pi_k (1-\pi_k))^2 + n_k \pi_k (1-\pi_k)(1-6\pi_k (1-\pi_k))\right) \end{aligned} \tag{A3}$$

Hence:

$$E\left[\frac{1}{K}\sum_{k=1}^{K} u_k^2\right] = E\left[\frac{1}{K}\sum_{k=1}^{K} E[u_k^2 \mid n_k, \pi_k]\right] = \frac{1}{K}\sum_{k=1}^{K} E\left[\frac{\pi_k (1-\pi_k)}{n_k}\right] = E\left[\frac{\pi_k (1-\pi_k)}{n_k}\right]. \tag{A4}$$

Independence implies that:

$$Var\left(\frac{1}{K}\sum_{k=1}^{K} u_k^2\right) = \frac{1}{K^2}\sum_{k=1}^{K} Var(u_k^2). \tag{A5}$$

Using the Law of Iterated Expectations and combining results, we can then write:

$$Var(u_k^2) = E[u_k^4] - (E[u_k^2])^2$$
$$= E\left[\frac{3(\pi_k(1-\pi_k))^2}{n_k^2} + \frac{\pi_k(1-\pi_k)(1-6\pi_k(1-\pi_k))}{n_k^3}\right] - \left(E\left[\frac{\pi_k(1-\pi_k)}{n_k}\right]\right)^2 < \infty. \quad \text{(A6)}$$

Therefore, as $K \to \infty$,

$$Var\left(\frac{1}{K}\sum_{k=1}^{K} u_k^2\right) \to 0. \quad \text{(A7)}$$

Because mean-square convergence implies convergence in probability, it then follows that:

$$\frac{1}{K}\sum_{k=1}^{K} u_k^2 \xrightarrow{p} E\left[\frac{\pi_k(1-\pi_k)}{n_k}\right]. \quad \text{(A8)}$$

B. Derivation of equation (10)

The model of interest is given by equation (1):

$$w_k = \beta\pi_k + \sum_h \alpha_h z_{kh} + \varepsilon_k, \quad \text{(A9)}$$

where $E\left[\varepsilon_k \middle| \pi_k, Z_k\right] = 0$, $\varepsilon_k \sim IID$, and $k = 1,...,K$. By definition, $p_k = \pi_k + u_k$, with $u_k \perp (Z_k, w_k)$ and $E(u_k \pi_k \mid Z_k) = 0$, with Z_k being the vector of explanatory variables in (A9) with dimension H. The relevant coefficient from the OLS regression when π_k is replaced with p_k is:

$$\hat{\beta}_p = \frac{\hat{C}ov(w_k, r_k)}{\hat{V}ar(r_k)} \quad \text{(A10)}$$

where r_k are the residuals from a regression of p_k on Z_k. Let ρ_k be the residuals from the unfeasible regression of π_k on Z_k, and define:

$$\hat{\beta} = \frac{\hat{C}ov(w_k, \rho_k)}{\hat{V}ar(\rho_k)}. \quad \text{(A11)}$$

Note that:

$$r = M(\pi + u)$$
$$\rho = M\pi \quad \text{(A12)}$$

where $M \equiv [I - Z(Z'Z)^{-1}Z']$, Z is the matrix of regressors, and π and u are $K \times 1$ vectors. We can then rewrite:

$$\hat{\beta}_p = ((\pi + u)'M(\pi + u))^{-1}((\pi + u)'Mw) \tag{A13}$$
$$\hat{\beta} = (\pi'M\pi)^{-1}(\pi'Mw)$$

Note that:

$$E[(\pi + u)'Mw] = E[\pi'Mw]. \tag{A14a}$$
$$E[((\pi + u)'M(\pi + u))] = E(\pi'M\pi) + E(u'Mu) + 2E(\pi'Mu) = E(\pi'M\pi) + E(u'Mu). \tag{A14b}$$

Equation (A14a) implies that $\hat{C}ov(w_k, r_k) = \hat{C}ov(w_k, \rho_k)$. Further, for any scalar random variable *a,* we note that $E(a) = E(tr\,(a)) = tr\,[E(a)]$, where $tr\,(\cdot)$ is the trace of the matrix in parentheses. It then follows that:

$$E[u'Mu] = tr\,E(u'Mu) = tr[E(Muu')] = (K - H)Var(u_k). \tag{A15}$$

Given equations (A14b) and (A15), we can then write an approximately unbiased estimator of $Var(\rho_k)$ as:

$$\begin{aligned} \hat{V}ar(\rho_k) &= \frac{E[\pi'M\pi]}{K} = \frac{E[p'Mp] - E[u'Mu]}{K}. \\ &= \hat{V}ar(r_k) - \frac{(K - H)}{K}\hat{V}ar(u_k) \end{aligned} \tag{A16}$$

Substituting the various definitions in (A11) implies that a consistent estimator is given by:

$$\hat{\beta} = \frac{\hat{C}ov(w_k, r_k)}{\hat{V}ar(r_k) - \left(\frac{K - H}{K}\right)\hat{V}ar(u_k)}, \tag{A17}$$

where, as derived in the text, $\hat{V}ar(u_k)$ can be approximated by $(1 - \tau)\frac{\bar{p}(1 - \bar{p})}{\bar{n}}$. The estimator defined in (A17) is subject to finite sample bias because of the nonlinear transformation applied separately to the unbiased estimators of the numerator and denominator. This bias is removed by using the bootstrap method described in the text.

Data Appendix: Construction of samples and variables

***Canada*:**

The data are drawn from the 1971, 1981, 1986, 1991, 1996 and 2001 Canadian Census microdata files maintained by Statistics Canada. Each of these confidential data files represents a 20 percent sample of the Canadian population, except for the 1971 file which represents a 33.3 percent sample. Statistics Canada also provides Public Use Microdata Files (PUMFs) to Canadian post-secondary institutions and to other researchers. The public use samples represent a much smaller proportion of the Canadian population (e.g., a 2.7 percent sample in 2001). The analysis is restricted to men aged 18-64. A person is classified as an immigrant if he reports being a landed immigrant in the Canadian census, and is either a noncitizen or a naturalized Canadian citizen; all other persons are classified as natives. Unless otherwise noted, sampling weights are used in all calculations. While information on age, sex, marital status, mother tongue and relationship to the "householder" are asked of 100% of the population the rest of the census information is obtained on a stratified 20% sample using the additional questions on the long form questionnaire. Weights in the Census files are used to project the information gathered from the 20% sample to the entire population.

Definitions of education and experience: We use the Census variables *dgreer* indicating "highest degree, certificate and diploma" and *trnucr* indicating "trade or non-university certificate" for the 1981 to 2001 Censuses to classify workers into five education groups: high school dropouts; workers with either a high school diploma or a vocational degree; workers with both a high school and vocational degree or a post-secondary certificate or diploma below Bachelor's degree; Bachelor's degree holders; and post-graduate degree holders. The coding of the relevant variables changes across Censuses. For the 2001 Census these five groups are identified by i) *dgreer*=1 or 11; ii) *dgreer*=2 or (*dgreer*=3 and *trnucr* $\neq$ 5 and *trnucr* $\neq$ 7); iii) *dgreer*=4 or *dgreer*=5 or (*dgreer*=3 and *trnucr*=5 or 7); iv) *dgreer*=6; and v) *dgreer*=7, 8, 9 or 10. The highest degree variable in the 1971 Census only identifies university degree, certificate and diploma holders (and aggregates all others as "not applicable").

The 1971 Census does not contain similar information on degrees. We instead rely on information about years of grade school (*highgrad*), vocational training (*training*), and years of post-secondary education below university (*otheredu*) to create classifications comparable to later Census years. Our construction of the education categories in 1971 assumes that if a worker does not have a Bachelor's degree but has 2 or more years of post-secondary education below university level, that worker possesses a post-secondary certificate or diploma. We also assume that Canadians who have eleven or more years of grade school and were born in Newfoundland or Quebec Provinces are high school graduates. All other Canadian-born and all immigrant men need 12 or more years of grade school to be considered high school graduates. This assumption recognizes the existence of different schooling systems across provinces and assumes that a Canadian-born worker's entire grade school education is completed in the province where they were born.

Canadian censuses also provide detailed information on the number of years an individual attended grade school (the variable *hgradr* in the 2000 census), post-secondary education below university (*ps_otr*), and university (*ps_uvr*). We calculate the total years of schooling by adding these variables and define work experience as Age - Years of Education - 6. We restrict the analysis to persons who have between 1 and 40 years of experience. Workers are

classified into one of 8 experience groups. The experience groups are defined in five-year intervals (1-5 years of experience, 6-10, 11-15, 16-20, 21-25, 26-30, 31-35, and 36-40).

Counts of persons in education-experience groups: The counts are calculated in the sample of men who do not reside in collective households, worked at some point in the past year (i.e., have a positive value for weeks worked in the previous calendar year), are not enrolled in school, and are not in the armed forces during the reference week[39]. The 1986 census does not provide school attendance information so that the construction of the 1986 sample ignores the school enrollment restriction. Our results are not sensitive to the exclusion of this cross-section from the analysis.

Annual and weekly earnings: We use the sample of men who do not reside in collective households, reported positive weeks worked and hours worked (during the reference week), are not in the armed forces in the reference week, and report positive earnings (sum of *wages*, *farmi*, and *selfi* variables, using the variable names corresponding to the 2001 Census). The 1971 census reports weeks worked in the calendar year prior to the survey as a categorical variable. We impute weeks worked for each worker as follows: 7 weeks for 1 to 13 weeks, 20 for 14-26 weeks, 33 for 27-39 weeks, 44 for 40-48 weeks and 50.5 for 49-52 weeks. The average log weekly earnings for a particular education-experience cell is defined as the mean of log weekly earnings over all workers in the relevant population.

United States:

The data are drawn from the 1960, 1970, 1980, 1990, and 2000 Integrated Public Use Microdata Samples (IPUMS) of the U.S. Census. In the 1960 Census, the data extract forms a 1 percent sample of the population. In the 1970 Census, the extract forms a 3 percent sample (obtained by pooling the state, metropolitan area, and neighborhood Form 1 files). In 1980, 1990, and 2000, the data extracts form a 5 percent sample. The analysis is restricted to men aged 18-64. A person is classified as an immigrant if he was born abroad and is either a non-citizen or a naturalized citizen; all other persons are classified as natives. Unless otherwise noted, sampling weights are used in all calculations. According to the Bureau of the Census, "the PUMS weight is a function of the full census sample weight and the PUMS sample design."[40]

Definition of education and experience: We use the IPUMS variables *educrec* to first classify workers into four education groups: high school dropouts (*educrec* <= 6), high school graduates (*educrec* = 7), persons with some college (*educrec* = 8), college graduates (*educrec* = 9). The college graduate sample is split into workers with 16 years of schooling or with post-graduate degrees using the variables *higrade* (in 1960-1980) and *educ99* (1990-2000). We assume that high school dropouts enter the labor market at age 17, high school graduates at age 19, persons with some college at age 21, college graduates at age 23, and workers with post-graduate degrees at age 25, and define work experience as the worker's age at the time of the survey minus the assumed age of entry into the labor market. We restrict the analysis to persons who have between 1 and 40 years of experience. Workers are classified into one of 8 experience groups, defined in five-year intervals.

---

[39] Note that the definition of the supply variables ignores the interesting issues that arise if immigration also influences native enrollment decisions or hours worked.

[40] The description of the sampling weights is found in: http://www.census.gov/prod/cen2000/doc/pums.pdf.

Counts of persons in education-experience groups: The counts are calculated in the sample of men who do not reside in group quarters, worked at some point in the past year (i.e., have a positive value for weeks worked in the period calendar year), are not enrolled in school, and are not in the military during the survey week.

Annual and weekly earnings: We use the sample of men who do not reside in group quarters, reported positive weeks worked and hours worked (last week's hours in 1960 and 1970; usual hours in 1980 through 2000), are not in the military in the reference week, and report positive earnings. Our measure of earnings is the sum of the IPUMS variables *incwage* and *incbusfm* in 1960, the sum of *incearn*, *incbus*, and *incfarm* in 1970 and 1980, and is defined by *incearn* in 1990-2000. In the 1960, 1970, and 1980 Censuses, the top coded annual salary is multiplied by 1.5. In the 1960 and 1970 Censuses, weeks worked in the calendar year prior to the survey are reported as a categorical variable. We imputed weeks worked for each worker as follows: 6.5 weeks for 13 weeks or less, 20 for 14-26 weeks, 33 for 27-39 weeks, 43.5 for 40-47 weeks, 48.5 for 48-49 weeks, and 51 for 50-52 weeks. The average log weekly earnings for a particular education-experience cell is defined as the mean of log weekly earnings over all workers in the relevant population.

References

Aigner, Dennis J. 1973. Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics* 1, no. 1:49-59.

Altonji, Joseph G., and David Card. 1991. The effects of immigration on the labor market outcomes of less-skilled natives. In *Immigration, Trade, and the Labor Market*, ed. John M. Abowd and Richard B. Freeman. Chicago: University of Chicago Press.

Angrist, Joshua D., and Alan B. Krueger. 1995. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business and Economic Statistics* 13, no. 2:225-235.

------ . 1999. Empirical strategies in labor economics, in *Handbook of Labor Economics*, Volume 3A, ed. Orley Ashenfelter and David Card. Amsterdam: Elsevier Science.

Aydemir, Abdurrahman, and George J. Borjas. 2007. A comparative analysis of the labor market impact of international migration: Canada, Mexico, and the United States. *Journal of the European Economic Association* 5, no. 4:663-708.

Bohn, Sarah, and Seth Sanders. 2005. Refining the estimation of immigration's labor market effects. Unpublished manuscript. Department of Economics, University of Maryland..

Bonin, Holger. 2005. Is the demand curve really downward sloping? Unpublished manuscript, IZA, Bonn, Germany.

Borjas, George J. 1987. Immigrants, minorities, and labor market competition. *Industrial and Labor Relations Review* 40, no. 3:382-392.

Borjas, George J. 2001. Does immigration grease the wheels of the labor market? *Brookings Papers on Economic Activity*, no. 1:69-119.

Borjas, George J. 2003. The labor demand curve *is* downward sloping: reexamining the impact of immigration on the labor market. *Quarterly Journal of Economics* 118, no. 4:1335-1374.

Borjas, George J. 2006. Native internal migration and the labor market impact of immigration. *Journal of Human Resources* 41, no. 2:221-258.

Borjas, George J., Richard B. Freeman and Lawrence F. Katz. 1997. "How much do immigration and trade affect labor market outcomes?" *Brookings Papers on Economic Activity*, no. 1:1-67.

Bound, John, Charles C. Brown, and Nancy Mathiowetz. 2001. Measurement error in survey data. In *Handbook of Econometrics*, ed. James J. Heckman and Edward E. Leamer. Amsterdam: Elsevier Science.

Bound, John, and Alan B. Krueger. 1991. The extent of measurement error in longitudinal data: do two wrongs make a right? *Journal of Labor Economics* 9, no. 1:1-24.

Butcher, Kristin F., and David Card. 1991. Immigration and wages: evidence from the 1980's. *American Economic Review Papers and Proceedings* 81, no. 2:292-296.

Cadena, Brian. 2010. Low-skilled immigration inflows and native competition. Unpublished Manuscript. University of Colorado, Boulder.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Card, David. 1990. The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review* 43, no. 2:245-257.

Card, David. 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, no. 1:22-64.

Deaton, Angus. 1985. Panel data from time series of cross-sections. *Journal of Econometrics* 30, nos. 1-2:109-126.

Filer, Randall K. 1992. The impact of immigrant arrivals on migratory patterns of native workers. In *Immigration and the Work Force: Economic Consequences for the United States and Source Areas*, ed. George J. Borjas and Richard B. Freeman. Chicago: University of Chicago Press.

Freeman, Richard B. 1984. Longitudinal analyses of the effect of trade unions. *Journal of Labor Economics* 2, no. 1:1-26.

Frey, William. 1995. Immigration impacts on internal migration of the poor: 1990 Census evidence for U.S. states. *International Journal of Population Geography* 1: 51-67.

Friedberg, Rachel M., and Jennifer Hunt. 1995. The impact of immigration on host country wages, employment and growth. *Journal of Economic Perspectives* 9, no. 2:23-44.

Garber, Steven, and Steven Keppler. 1980. Extending the classical normal errors-in-variables model. *Econometrica* 48, no. 6:1541-1546.

Greene, William H. 1993. *Econometric Analysis*, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

Griliches, Zvi, and William M. Mason. 1972. Education, income and ability. *Journal of Political Economy* 80, no. 3, part 2:74-103.

Griliches, Zvi, and Jerry Hausman. 1986. Errors in variables in panel data. *Journal of*

Econometrics 36, no. 1:93-118.

Grossman, Jean Baldwin. 1982. The substitutability of natives and immigrants in production. *Review of Economics and Statistics* 54, no. 4:596-603.

Gurak, Douglas T., and Mary M. Kritz. 2000. The interstate migration of U.S. immigrants: individual and contextual determinants. *Social Forces* 78, no. 3:1017-39.

Hartog, Joop, and Aslan Zorlu. 2005. "The effect of immigration on wages in three European countries. *Journal of Population Economics* 18, no.1:113-151.

Horowitz, Joel L. 2001. The bootstrap in econometrics. In *Handbook of Econometrics*, Vol. 5, ed. James J. Heckman and Edward E. Learner. Amsterdam: Elsevier Science.

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee. 1982. *Introduction to the theory and practice of econometrics*. New York: John Wiley & Sons.

Kane, Thomas J., Cecilia E. Rouse, and Douglas Staiger. 1999. "Estimating returns to schooling when schooling is misreported. Working Paper No. 7325, National Bureau of Economic Research, Cambridge, MA.

Kmenta, Jan. 1997. *Elements of econometrics*, *Second Edition*. Ann Arbor, MI: The University of Michigan Press.

Kritz, Mary M., and Douglas T. Gurak. 2001. The impact of immigration on the internal migration of natives and immigrants. *Demography* 38, no. 1:133-45.

LaLonde, Robert J., and Robert H. Topel. 1991. Labor market adjustments to increased immigration. In *Immigration, Trade, and the Labor Market*, ed. John M. Abowd and Richard B. Freeman. Chicago: University of Chicago Press.

Levi, Maurice D. 1973. Errors in the variables bias in the presence of correctly measured variables. *Econometrica* 41, no. 5:985-986.

Lewis, Ethan. 2005. Immigration, skill mix, and the choice of technique. Working Paper No. 05-08, Federal Reserve Bank of Philadelphia.

Longhi, Simonetta, Peter Nijkamp, and Jacques Poot. 2005. "A meta-analytic assessment of the effect of immigration on wages." *Journal of Economic Surveys* 19, no. 3:451-477.

Maddala, G.S. 1992. *Introduction to econometrics, Second Edition*. Englewood Cliffs, NJ: Prentice-Hall.

McKinnish, Terra. 2008. Panel data models and transitory fluctuations in the explanatory variable. *Advances in Econometrics* 21:335-358.

Mishra, Prachi. 2007. Emigration and wages in source countries: evidence from Mexico. *Journal of Development Economics* 82, no. 1:180-199.

Paggiaro, Adriano, and Nicola Torelli. 2004. The effect of classification errors in survival data analysis. *Statistical Methods and Applications* 13, no. 2:213-225

Paxson, Christina, and Jane Waldfogel. 2002. Work, welfare, and child maltreatment. *Journal of Labor Economics* 20, no. 3:435-474.

Pischke, Jörn-Steffen, and Johannes Velling. 1997. Employment effects of immigration to Germany: an analysis based on local labor markets. *Review of Economics and Statistics* 79, no. 4:594-604.

Richardson, David H., and De-Min Wu. 1970. Least squares and grouping method estimators in the errors in variables model. *Journal of the American Statistical Association* 65, no. 330:724-748.

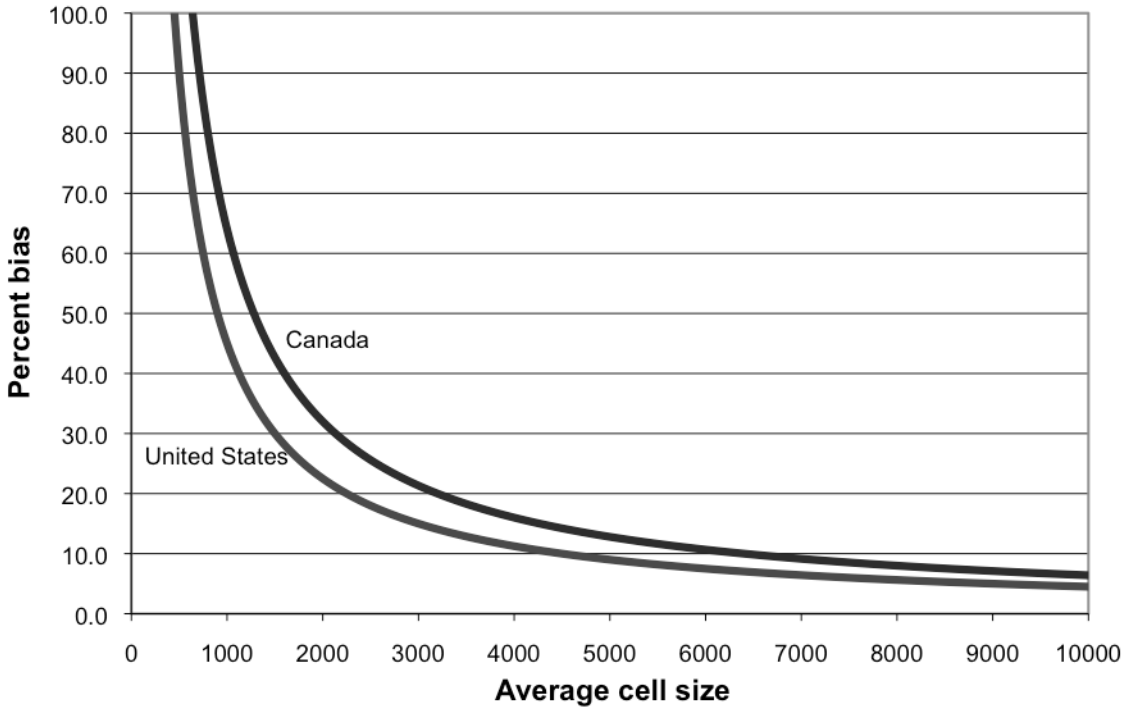
Saiz, Albert. 2003. Room in the kitchen for the melting pot: immigration and rental prices. *Review of Economics and Statistics* 85, no. 3: 502-521.

Schoeni, Robert F. 1997. The effect of immigrants on the employment and wages of native workers: evidence from the 1970s and 1980s. Working Paper No. DRU-1408-IF, The Rand Corporation, Santa Monica, CA.

Smith, James P., and Barry Edmonston, ed. 1997. *The new Americans: economic, demographic, and fiscal effects of immigration*. Washington, D.C.: National Academy Press.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20, no. 4:518-529.

Figure 1. Predicted percent bias on estimated wage impact of immigration in national labor market



Note: The simulation for Canada assumes that the mean immigrant share is 0.2; the variance of the immigrant share across national-level labor markets is 0.005; and the R^2 of the auxiliary regression is 0.95. The simulation for the United States assumes that the mean immigrant share is 0.1; the variance of the immigrant share across national-level labor markets is 0.004; and the R^2 of the auxiliary regression is 0.95.

Table 1. The observed distribution of the immigrant share, national-level analysis

	Statistics Canada file	5/100	PUMF	1/100	1/1000	1/10000
Canada:						
$\overline{n}$	30416.3	7000.7	3426.9	1399.8	139.9	14.4
$\overline{p}$	0.191	0.191	0.191	0.191	0.191	0.191
σ_p^2	0.0050	0.0050	0.0051	0.0051	0.0064	0.0194
10^{th} percentile	0.123	0.123	0.122	0.122	0.112	0.001
50^{th} percentile	0.229	0.229	0.228	0.228	0.220	0.193
90^{th} percentile	0.366	0.365	0.365	0.365	0.388	0.496
United States						
$\overline{n}$	---	47564.3	---	11746.0	1174.6	117.4
$\overline{p}$	---	0.077	---	0.077	0.077	0.077
σ_p^2	---	0.0037	---	0.0037	0.0037	0.0044
10^{th} percentile	---	0.035	---	0.035	0.034	0.021
50^{th} percentile	---	0.070	---	0.070	0.071	0.070
90^{th} percentile	---	0.152	---	0.152	0.162	0.188

Note: The variable $\overline{n}$ gives the average number of observations in the education-experience-year cell used to calculate the mean immigrant share; $\overline{p}$ gives the mean immigrant share across cells; and σ_p^2 gives the variance of the immigrant share across cells. All statistics reported in the table, except those referring to the Statistics Canada file and the 5/100 U.S. Census, are averages across 500 replications of random samples at the given sampling rate. The analysis of the Canadian labor market has 240 cells; the analysis of the U.S. labor market has 200 cells.

Table 2. Estimated wage impact of immigration, national-level analysis - Canada

	Stat. Can.	5/100	PUMF	1/100	1/1000	1/10000
Canada:						
1. $\hat{\beta}$	-0.507	-0.468	-0.403	-0.342	-0.076	-0.011
2. Standard error of $\hat{\beta}$	0.202	0.196	0.189	0.180	0.191	0.200
3. Standard deviation of $\hat{\beta}$	---	0.056	0.099	0.119	0.174	0.174
4. Fraction $\hat{\beta}$ significant at:						
1% level	---	0.308	0.232	0.178	0.018	0.008
5% level	---	0.912	0.624	0.458	0.064	0.048
10% level	---	0.988	0.820	0.644	0.108	0.074
5. R^2 of auxiliary regression	0.967	0.965	0.960	0.953	0.845	0.590
6. β^*	---	-0.505	-0.501	-0.499	-0.466	-0.384
7. Standard error of β^*	---	0.209	0.226	0.241	0.485	1.475
8. Standard deviation of β^*	---	0.049	0.093	0.126	0.405	1.353
Corrected coefficients:						
9. Back-of-the-envelope	-0.520	-0.531	-0.524	-0.638	1.174	0.044
10. Standard deviation of row 9	---	0.064	0.132	0.241	15.647	1.652
11. BBE method	-0.530	-0.539	-0.466	-0.680	-0.599	0.308
12. USSIV method	-0.519	-0.515	-0.520	-0.525	0.482	-0.486
13. Standard deviation of row 12	0.034	0.010	0.211	0.304	18.460	25.475

Note: The coefficient $\hat{\beta}$ gives the estimated wage impact of immigration; β^* gives the coefficient when the observed immigrant share is replaced by the immigrant share calculated from the largest file (i.e., the Statistics Canada file). The R^2 of the auxiliary regression gives the multiple correlation of the regression of the immigrant share on all other explanatory variables in the model. The corrected coefficients use the methods described in the text to net out the impact of sampling error on $\hat{\beta}$. All statistics reported in the table, except those referring to the Statistics Canada file, are averages across 500 replications of random samples at the given sampling rate. The analysis of the Canadian labor market has 240 cells.

Table 3. Estimated wage impact of immigration, national-level analysis – United States

	5/100	1/100	1/1000	1/10000
1. $\hat{\beta}$	-0.489	-0.476	-0.347	-0.082
2. Standard error of $\hat{\beta}$	0.223	0.225	0.247	0.279
3. Standard deviation of $\hat{\beta}$	---	0.056	0.162	0.227
4. Fraction $\hat{\beta}$ significant at:				
1% level	---	.062	.046	.006
5% level	---	.706	.212	.042
10% level	---	.962	.368	.072
5. R^2 of auxiliary regression	0.974	0.973	0.964	0.883
6. β^*	---	-0.488	-0.497	-0.498
7. Standard error of β^*	---	0.228	0.291	0.631
8. Standard deviation of β^*	---	0.045	0.171	0.534
Corrected coefficients:				
9. Back-of-the-envelope	-0.496	-0.506	-0.642	5.794
10. Standard deviation of row 9		0.060	0.320	89.464
11. BBE method	-0.529	-0.446	-0.035	0.089
12. USSIV method	-0.496	-0.496	-0.503	-5.420
13. Standard deviation of row 12	0.040	0.090	0.371	48.076

Note: The coefficient $\hat{\beta}$ gives the estimated wage impact of immigration; β^* gives the coefficient when the observed immigrant share is replaced by the immigrant share calculated from the largest file (i.e., rhe 5/100 U.S. Census). The R^2 of the auxiliary regression gives the multiple correlation of the regression of the immigrant share on all other explanatory variables in the model. The corrected coefficients use the methods described in the text to net out the impact of sampling error on $\hat{\beta}$. All statistics reported in the table, except those referring to the 5/100 U.S. Census, are averages across 500 replications of random samples at the given sampling rate. The analysis of the U.S. labor market has 200 cells.

Table 4. The observed distribution of the immigrant share, metropolitan area analysis

	Statistics Canada file	5/100	PUMF	1/100
Canada:				
$\overline{n}$	659.5	165.1	83.9	34.1
$\overline{p}$	0.233	0.233	0.233	0.233
σ_p^2	0.0227	0.0234	0.0245	0.0274
10^{th} percentile	0.022	0.002	0.000	0.000
50^{th} percentile	0.178	0.175	0.169	0.157
90^{th} percentile	0.407	0.427	0.447	0.482
United States:				
$\overline{n}$	---	174.4	---	35.7
$\overline{p}$	---	0.103	---	0.103
σ_p^2	---	0.0137	---	0.0152
10^{th} percentile	---	0.009	---	0.000
50^{th} percentile	---	0.061	---	0.000
90^{th} percentile	---	0.247	---	0.237

Note: The variable $\overline{n}$ gives the average number of observations in the city-education-experience-year cell used to calculate the mean immigrant share; $\overline{p}$ gives the mean immigrant share across cells; and σ_p^2 gives the variance of the immigrant share across cells. All statistics reported in the table, except those referring to the Statistics Canada file and the 5/100 U.S. Census, are averages across 500 replications of random samples at the given sampling rate. The analysis of the Statistics Canada file has 5,360 cells; the analysis of the 5/100 U.S. file has 31,472 cells.

Table 5. Estimated wage impact of immigration, metropolitan area analysis

	Statistics Canada	5/100	PUMF	1/100
Canada:				
1. $\hat{\beta}$	-0.053	-0.022	-0.012	-0.004
2. Standard error of $\hat{\beta}$	0.037	0.039	0.040	0.042
3. Standard deviation of $\hat{\beta}$	---	0.032	0.036	0.039
4. R^2 of auxiliary regression	0.982	0.959	0.929	0.864
Using "large sample" share				
5. β^*	---	-0.053	-0.055	-0.049
6. Standard error of β^*	---	0.060	0.083	0.127
7. Standard deviation of β^*	---	0.045	0.069	0.115
Corrected coefficients:				
8. Back-of-the-envelope	-0.112	0.328	0.065	0.009
9. Standard deviation of row 8	---	0.921	0.196	0.099
10. BBE method	-0.085	-0.015	-0.122	0.086
11. USSIV method	-0.094	-0.076	-0.096	0.962
12. Standard deviation of row 11	0.067	0.254	0.590	26.536
United States:				
1. $\hat{\beta}$	---	-0.050	---	-0.022
2. Standard error of $\hat{\beta}$	---	0.023	---	0.023
3. Standard deviation of $\hat{\beta}$	---	---	---	0.019
4. R^2 of auxiliary regression	---	0.948	---	0.896
Using "large sample" share				
5. β^*	---	---	---	-0.033
6. Standard error of β^*	---	---	---	0.064
7. Standard deviation of β^*	---	---	---	0.045
Corrected coefficients:				
8. Back-of-the-envelope	---	-0.170	---	0.036
9. Standard deviation of row 8	---	---	---	0.031
10. BBE method	---	-0.096	---	-0.082
11. USSIV method	---	-0.072	---	-0.068
12. Standard deviation of row 11	---	0.026	---	0.113

Note: The coefficient $\hat{\beta}$ gives the estimated wage impact of immigration; β^* gives the coefficient when the observed immigrant share is replaced by the immigrant share calculated from the largest file (i.e., the Statistics Canada file or the 5/100 U.S. Census). The R^2 of the auxiliary regression gives the multiple correlation of the regression of the immigrant share on all other explanatory variables in the model. The corrected coefficients use the methods described in the text to net out the impact of sampling error on $\hat{\beta}$. All statistics reported in the table, except those referring to the Statistics Canada file and the 5/100 U.S. Census, are averages across 500 replications of random samples at the given sampling rate. The analysis of the Statistics Canada file has 5,360 cells; the analysis of the 5/100 U.S. file has 31,472 cells.

Table 6. Sensitivity of first-stage coefficient in IV regression model

	Stat. Can.	5/100	PUMF	1/100	1/1000	1/10000
Canada:						
National level						
$\hat{\delta}$	0.258	0.207	0.155	0.054	-0.175	-0.224
Standard error	0.085	0.089	0.093	0.100	0.102	0.111
Pr($F > 10$)	0.000	0.042	0.016	0.002	0.078	0.158
$\hat{\delta}(p_t, \pi^*_{t-1})$	---	0.256	0.258	0.261	0.245	0.206
$\hat{\delta}(\pi^*_t, p_{t-1})$	---	0.231	0.201	0.149	0.029	0.003
Metropolitan area						
$\hat{\delta}$	0.121	-0.081	-0.130	-0.188	-0.234	-0.304
Standard error	0.026	0.022	0.021	0.021	0.039	0.434
Pr($F > 10$)	1.000	0.762	1.000	1.000	1.000	0.006
$\hat{\delta}(p_t, \pi^*_{t-1})$	---	0.123	0.124	0.133	0.214	0.342
$\hat{\delta}(\pi^*_t, p_{t-1})$	---	0.049	0.026	0.012	0.003	0.001
United States						
National level						
$\hat{\delta}$	---	0.464	---	0.433	0.165	-0.135
Standard error	---	0.218	---	0.219	0.197	0.134
Pr($F > 10$)	---	0.000	---	0.000	0.004	0.028
$\hat{\delta}(p_t, \pi^*_{t-1})$	---	---	---	0.464	0.457	0.453
$\hat{\delta}(\pi^*_t, p_{t-1})$	---	---	---	0.445	0.266	0.054
Metropolitan area						
$\hat{\delta}$	---	-0.108	---	-0.371	---	---
Standard error	---	0.022	---	0.017	---	---
Pr($F > 10$)	---	1.000	---	1.000	---	---
$\hat{\delta}(p_t, \pi^*_{t-1})$	---	---	---	-0.089	---	---
$\hat{\delta}(\pi^*_t, p_{t-1})$	---	---	---	-0.033	---	---

Note: The coefficient $\hat{\delta}$ and "standard error" give the estimated coefficient and standard error from the regression of the immigrant share on the lagged immigrant share; Pr($F > 10$) gives the probability that the F-statistic associated with this coefficient exceeds 10 under the null that the population coefficient equals zero; $\hat{\delta}(p_t, \pi^*_{t-1})$ is the coefficient from the regression of the observed immigrant share on the lagged "true" share calculated in the largest available sample; and $\hat{\delta}(\pi^*_t, p_{t-1})$ gives the coefficient from the regression of the true immigrant share on the lagged observed share. All statistics, except those referring to the Statistics Canada file and the 5/100 U.S. Census, are averages across 500 replications of random samples at the given sampling rate. The analysis of the Statistics Canada file has 160 cells in the national-level analysis and 4,288 cells in the metropolitan area analysis. The analysis of the 5/100 U.S. Census file has 160 cells at the national-level and 17,510 cells at the metropolitan area level.