Attenuation Bias in Measuring the Wage Impact of Immigration

Abdurrahman Aydemir and George J. Borjas
Statistics Canada and Harvard University

July 2006

Attenuation Bias in Measuring the Wage Impact of Immigration

Abdurrahman Aydemir and George J. Borjas

ABSTRACT

Although economic theory predicts that there should be an inverse relation between relative wages and immigrant-induced supply shifts, many empirical studies have found it difficult to document such effects. We argue that much of the weak evidence may be due to sampling error in the most commonly used measure of the immigrant supply shift: the fraction of the workforce that is foreign-born. Because the immigrant share is a proportion, its sampling error can be easily derived from the properties of the hypergeometric distribution. Sampling error plays a disproportionately large role because the typical study is longitudinal, investigating how wages adjust as immigrants enter a particular labor market. After controlling for permanent factors that determine wages in specific labor markets, there is little variation remaining in the immigrant share. Using data for both the Canadian and U.S. labor markets, we find that there is indeed significant measurement error in this measure of immigrant supply shifts, and that correcting for the attenuation bias can easily more than triple existing estimates of the wage impact of immigration.

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 immigrant-induced labor supply shift, at least in the short run. In particular, higher levels of immigration should lower the wage of competing workers and increase the wage of complementary workers.

Despite the common-sense intuition behind these predictions, the economics literature has—at least until recently—found it difficult to document the inverse relation between wages and immigrant-induced supply shifts. Much of the literature attempts to estimate 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 studies, there is some consensus that the estimates 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 a Senior Economist at Statistics Canada; Dr. Borjas is a Professor of Economics and Social Policy at the Kennedy School of Government, Harvard University, and a Research Associate at the National Bureau of Economic Research. We are grateful to Alberto Abadie, Sue Dynarski, Richard Freeman, Daniel Hamermesh, Larry Katz, and Douglas Staiger for very helpful discussions and comments. This paper represents the views of the authors and does not necessarily reflect the opinion of Statistics Canada.

[1] Representative studies include Altonji and Card (1991), Borjas (1987), Borjas, Freeman, and Katz (1997), Card (1991, 2001), Grossman (1982), LaLonde and Topel (1991), Pischke and Velling (1997), and Schoeni (1997). Friedberg and Hunt (1995) and Smith and Edmonston (1997) 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 only be measurable 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 local labor market studies, the national labor market evidence indicated that wage growth was strongly and inversely related to immigrant-induced supply increases.

A number of papers have already replicated the national-level Borjas framework, 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 (2003) 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 immigrant-induced supply shifts. In contrast, Bohn and Sanders (2005) use publicly available Canadian data and report near-zero factor price elasticities for the Canadian labor market.

---

[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. 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. 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.

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 great deal 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.

Measurement error plays a central role in these studies because of the longitudinal nature of the empirical exercise that is conducted. Immigration is often measured by the "immigrant share," the fraction of the workforce in a particular labor market that is foreign-born. The analyst then typically examines the relation between the wage and the immigrant share *within* a particular labor market. To net out market-specific wage effects, the study 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 the sampling error in the immigrant share to play a disproportionate large role. As a result, even very small amounts of sampling error get magnified and easily dominate the remaining variation in the immigrant share.

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 (e.g., how the sampling error depends on the sample size used to calculate the immigrant share) 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 quantitative 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.[3] 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 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 (e.g., the post-1980 5 percent samples) reveals significant effects, but the effects become exponentially weaker as the analyst calculates the immigrant share variable on progressively smaller samples.

In short, there is significant measurement error in the variable that has been traditionally used to measure the immigrant supply shock in a labor market, whether at the national level for a skill group or at the local level. The smaller the sample used to calculate the fraction of the workforce in the market that is foreign-born, the smaller the resulting wage impact of immigration. Our analysis indicates that correcting for the attenuation bias caused by sampling error can easily more than triple existing estimates of the wage impact of immigration.

[3] These confidential files are the largest available micro data files in Canada that provide information on citizenship, immigration, schooling, labor market activities, and earnings.

II. Framework

Suppose we are interested in estimating the wage impact of immigration by looking at wage variation in labor markets. The labor markets may be defined in terms of skills, geographic regions, and 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:

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

where w_k gives the log wage in labor market k ($k = 1,\ldots,K$); 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 X 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.

An important characteristic of this type of empirical exercise is that the analyst typically calculates the immigrant share from the microdata available for labor market k. For example, the 2000 U.S. Census contains many observations on workers who are high school dropouts and are aged 30-34 in 2000. The analyst would then calculate the immigrant share for this particular labor market (defined by education, age, and time) from these individual observations. This type of calculation inevitably introduces sampling error in the key independent variable in equation (1), and introduces the possibility that the coefficient β may be inconsistently estimated.

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 workers by immigrant status.[4] 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 the 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 $\text{Var}(u_k) = \pi_k(1 - \pi_k)/n_k$. Census sampling, however, is typically 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) used to generate the Census sample (e.g., a 1/1000 sample). We can then approximate the variance of the error term in (2) by $\text{Var}(u_k) = (1- \tau)\, \pi_k (1 - \pi_k)/n_k$. Note that for very small sampling rates, the variance of the

[4] 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). It is well known that the skill distribution of immigrant workers in the United States varies across cities and regions, so these types of regressions are unlikely to capture the true wage impact of immigration.

sampling error has a simple binomial structure.[5] As we will see below, the statistical properties of the sampling error have important implications for the size of the attenuation bias in estimates of the wage impact of immigration. They also provide relatively simple ways for correcting the estimates for the impact of measurement error. Finally, note that u_k and π_k are mean-independent; this implies $\text{Cov}(\pi_k, u_k) = 0$.

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:[6]

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

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 variance of the observed immigrant share that remains unexplained after controlling for all other variables in the regression model.

---

[5] 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.

[6] Maddala (1992, pp. 451-454) presents a particularly simple derivation of equation (3) when the regression has two explanatory variables; see also Levin (1973) and Garber and Keppler (1980). Although most of the discussions on measurement error focus on the bias of the coefficient, it is also the case that measurement error biases the standard errors. Bloch (1978) shows that (in the one explanatory variable case) measurement error biases the standard error of the mismeasured variable, but the direction of the bias is ambiguous. Nevertheless, the *t*-statistic is biased towards zero. Bound, Brown and Mathiowetz (2001) provide an excellent survey of the econometric problems associated with measurement error.

As noted above, the typical study in the literature often pools data on particular labor markets over time and then adds fixed effects that net out persistent wage effects in labor market *k* as well as period effects. This type of regression model, of course, is equivalent to differencing the 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 important role in the estimation.

It can be shown that the probability limit of the average of the square of error terms in (3) is:

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

where the expectation in (4) is taken across the *K* labor markets. Combining results, we can write the expression for the probability limit of the coefficient measuring the wage impact of immigration in the presence of sampling error as:

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

Equation (5) imposes an important restriction on the magnitude of the measurement error. In particular, note that 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

measurement 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 information about the value of the true parameter from observed data. We show below that this restriction is often violated by the immigrant share calculated in relatively small samples.

These violations may arise for two reasons. First, any calculation of the expected sampling error in (5) will require 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, the derivation of equation (5) assumes 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. It is well known that if immigrants (and natives) are systematically misclassified, the measurement error in the observed immigrant share will be correlated with the true value of the immigrant share (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 will be convenient to present an approximation to equation (5) that can be used to get a back-of-the-envelope estimate of the quantitative impact 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:

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

where $\overline{n}$ (= n_T/K) is the 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 $\overline{p}(1-\overline{p})$, where $\overline{p}$ is the average observed immigrant share across the K labor markets.[7] This approximation implies that we can rewrite equation (5) as:

$$
\text{plim}\ \hat{\beta} \approx \beta\left(1-(1-\tau)\frac{\overline{p}(1-\overline{p})/\overline{n}}{(1-R^2)\sigma_p^2}\right) \tag{7}
$$

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 (and not surprisingly), 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

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

the order of 0.95. Equation (7) implies that the percent bias generated by sampling error is given by:

$$\text{(8)} \qquad \text{Percent bias} \approx (1-\tau)\frac{\overline{p}(1-\overline{p})/\overline{n}}{(1-R^2)\sigma_p^2}.$$

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 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.[8]

The figure also reports the results of a similar simulation 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

[8] Under the maintained assumption that $\overline{p} = 0.1$, $\sigma_p^2 = 0.004$, and $R^2 = 0.95$, the bias cannot be calculated if the average cell size is less than 450. The implied amount of measurement error would then be larger than the unexplained variance in the immigrant share.

the longitudinal nature of the exercise removes much of the identifying variation in the immigrant share—implies a sizable 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.

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. We will show below, 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 numerically important.

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 microdata 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 data 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 Sample (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.[9] 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.[10]

---

[9] 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.

[10] 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.

A. National Labor Market

As noted earlier, Borjas (2003) suggests that the wage impact of immigration can 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

We classify workers in both the Canadian and U.S. labor markets into five distinct education groups: (1) high school dropouts; (2) high school graduates; (3) workers who have some college; (4) college graduates; and (5) workers with post-graduate education. We classify workers into a particular years-of-experience cohort by using potential years of experience, roughly defined by Age – Years of Education – 6. The analysis is restricted to persons who have between 1 and 40 years of experience. 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.

Our classification of the education and experience groups 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 can attenuate 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 files resident at Statistics Canada. These data indicate 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.[11]

The remaining columns of the top panel show what happens to this distribution as we consider progressively smaller samples of the Canadian workforce. In particular, we examine 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 publicly available in the PUMF) generate a relatively small sample size for the average cell even at the level of the *national* Canadian labor market.

---

[11] 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. 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 each cross-section to prevent the more recent cross-sections from contributing more to the estimation

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 will inevitably 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 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. As we showed in the previous section, the importance of sampling error in generating biased coefficients becomes exponentially greater as the average cell size declines, so that national-level studies of the labor market impact of immigration in Canada could be greatly affected by attenuation bias.

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

---

simply because each country's population increased over time. The evidence presented below is not sensitive to the choice of weights.

10th percentile cell has an immigrant share of 12.3 percent. Table 1 shows that the immigrant share in the 10th percentile cell declines as the sample size falls. In the 1/1000 replications, for example, the 10th 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 90th percentile cell in the Statistics Canada files has an immigrant share of 36.6 percent. In the 1/1000 replication, however, the 90th 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."

The data for the U.S. labor market—where we only have access to the publicly available 5 percent files in the post-1980 period and to smaller samples prior to 1980—tell the same story. As with our analysis of the Canadian data, we use these data 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 roughly 10 times larger than that of Canada, note that 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:

$$(9) \qquad w_{sxt} = \beta\, p_{sxt} + S + X + T + (S \times X) + (S \times T) + (R \times T) + \varepsilon_{sxt},$$

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. As noted above, the regressions weigh 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.

The top panel of 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.[12] 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

[12] It is easier to interpret this coefficient by converting it to a wage elasticity that gives the percent change in wages associated with a percent 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 2 can be interpreted as wage elasticities by multiplying the coefficient by approximately 0.6.

bias caused by sampling error could easily play a huge 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 the top panel of Table 2 document the impact of measurement error as we estimate the same regression model on progressively smaller samples.

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 already falls by 7.7 percent; the coefficient now equals –0.468 and has an average standard error of 0.196. 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. The results from this particular analysis are worth emphasizing because 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), and the average standard error is 0.189. In short, the typical researcher using what seems to be a large publicly available random sample of Canadian workers would conclude that immigration had a much smaller numerical impact on wages.[13] In fact, we can drive the estimate of β to zero by

[13] This is not idle speculation. Bohn and Sanders (2005), for example, attempt to 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 single 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

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 only 14 observations per cell and the average coefficient is -0.011, with an average standard error of 0.200.

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 the attenuation bias generated by 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 measurement 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 5-7 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, with a standard error of 0.202. If we used the correct immigrant share in the 1/100 replications the estimated coefficient β^* is -0.499, but 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, but 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

---

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.

across replications of 1.353. In sum, the results summarized in Table 2 provide compelling evidence that sampling error in the measure of the immigrant share can greatly attenuate the estimated wage impact of immigration.

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 consistent regression coefficients. It is important, therefore, to consider alternative methods of correcting the regression coefficients for the attenuation bias induced by sampling error. Equation (7) provides a very 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.[14] In particular, we can do 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 cell size. We calculated the corrected coefficient for each of the 500 replications (at each sampling rate). Row 8 of Table 2 reports the average corrected coefficient and row 9 reports the standard deviation across replications.

Alternatively, we can directly estimate the mean of the sampling error defined in equation (4) by using the available information on immigrant shares and cell size for the K cells in the analysis. More precisely, let:

[14] Some of the replications combine samples collected at different sampling rates. The sampling rate is set at 0.20 for the corrections in the Canada Statistics 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.

$$(10) \qquad E\left(\frac{\pi_k(1-\pi_k)}{n_k}\right)=\frac{\sum_k \lambda_k p_k(1-p_k)/n_k}{\sum_k \lambda_k},$$

where the weight λ_k gives the number of native workers in cell *k* and the sum of the weights is normalized to one in each cross-section. We calculate the expectation in (10) for each of the 500 replications (at each sampling rate). We then use this statistic to adjust the estimated coefficient $\hat{\beta}$ in each replication. Row 10 of the table reports the average corrected coefficient and row 11 reports the standard deviation. It is worth emphasizing that this calculation can generate imprecise results (particularly for small samples) because we are using the observed immigrant share p_k as an estimate of the true share π_k. If, for example, both the true immigrant share and cell size are relatively small, the observed immigrant share will likely be zero and this particular cell will not contribute to the calculation of the mean sampling error.

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.

Both methods of correction generate adjusted coefficients that typically approximate this "true" effect as long as the mean cell size is large, but are much less precise when the mean cell size declines. In the 5/100 replication, for example, both correction methods lead to adjusted coefficients of around -0.53. By the time we get to the PUMF sampling rate, however, the magnitude of the adjustment implied by the two methods of correction begins to diverge. At this

sampling rate, the inconsistent coefficient $\hat{\beta}$ is -0.403. The average adjusted coefficient is -0.52 if we use the back-of-the-envelope approach in equation (7), or -0.59 if we use the more complex approach in equation (10). Both adjusted coefficients are further off the mark if we move to the 1/100 replications. The estimates are -0.64 and -0.69, respectively, with very large standard deviations. Finally, if the cell size gets sufficiently small, as in the 1/1000 replication, both correction methods break 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. Put differently, at such low sampling rates, the data do not provide sufficient information to allow the analyst to recover the true wage impact of immigration.

It is worth emphasizing that although there is relatively little difference in the adjustments implied by the two corrections for large samples, the back-of-the-envelope approach in equation (7) provides better estimates of the true wage impact of immigration for medium-sized samples. The likely reason is that the use of cell-level information on the immigrant share introduces inaccuracies in the calculation of the mean binomial error that are "washed out" by simply using the mean immigrant share in the entire sample.

It is also of interest to compare these corrections to a more sophisticated regression-based approach. The hypergeometric distribution of the sampling error indicates that we have a great deal of information about the reliability of the immigrant share variable that can be explicitly introduced into the regression estimation. In particular, write the regression model as:

$$\mathbf{y} = \mathbf{X}\theta + \boldsymbol{\varepsilon}, \tag{11}$$

$$\mathbf{X} = \mathbf{X}^{*} + \mathbf{u}, \tag{12}$$

where $\mathbf{X}^*$ gives the "true" values of the regressors. Assuming that the error terms in the model are normally distributed, the error-corrected coefficient vector $\boldsymbol{\theta}$ can be estimated by $\hat{\boldsymbol{\theta}} = \mathbf{B}^{-1}\mathbf{X}'\mathrm{y}$, with $\mathbf{B} = \mathbf{X}'\mathbf{X} - \mathbf{S}$, where $\mathbf{S}$ is a matrix with diagonals equal to $K(1-\rho_h)s_h^2$ and zero elsewhere; ρ_h is the reliability index (defined as 1 – fraction of the total variance in the regressor that is due to error), and s_h^2 is the variance of regressor *h*. In our analysis, we only permit the immigrant share variable to be measured with error (so all the other diagonals are set to zero), and the extent of error is known because of the hypergeometric sampling distribution. It is relatively easy, therefore, to generate error-corrected regression coefficients.[15]

The estimated corrected regression coefficients are reported in row 9 of Table 2 (and the standard error is reported in row 10). The regression-based estimates are very similar to those estimated using the simpler corrections. Consider, for instance, the results obtained in the PUMF replication. The coefficient estimated in the regression that uses the mismeasured immigrant share variable is -0.403; the back-of-the-envelope correction in row 7 yields a predicted coefficient of -0.524; and the regression-based method also yields a prediction of -0.524.[16]

The bottom panel of Table 2 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:

---

[15] We calculate the reliability index ρ_h using the back-of-the-envelope approximation to the mean hypergeometric error in equation (7); see Judge et. al (1982, pp. 534-537) and Kmenta (1986, pp. 352-354) for details. We used the statistical command *eivreg* in STATA to compute the regression-based adjusted coefficients.

[16] It is also possible to use instrumental variables to correct for measurement error bias. We will discuss below the statistical 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*.

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 estimated wage impact of immigration to just under -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 2 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. Despite the fact that the immigrant share is calculated in samples that contain over 100 workers on average, that variable contains little valuable information that can be used in any empirical study of the wage impact of immigration.

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 4 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. Note, however, that the coefficient β^* is not precisely estimated. The average standard error in the 1/10000 replications is 0.631, and the standard deviation of this coefficient across replications is 0.534.

The remaining rows of the bottom panel of the table show what happens to the estimated wage impact of immigration when we use the various correction methods to adjust the inconsistent estimate for sampling error. The corrections conducted in the 1/100 and 1/1000 replications work reasonably well: the corrected coefficients lie between -0.5 and -0.6 when we use the back-of-the-envelope method in equation (7). Note, however, that none of the correction methods lead to sensible predictions in the 1/10000 replication because the predicted sampling error often exceeds the adjusted variance in the immigrant share.

B. Spatial Correlations

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 most of the spatial correlation 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.[17] 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.[18] It immediately follows that the number of observations per cell declines substantially once we move the unit of analysis to this level of geography.[19]

Table 3 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 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.

---

[17] 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.

[18] 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.

[19] 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.

As with our study of the distribution of the immigrant share in national-level labor markets, 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. Note, however, that 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:

$$\log w_{hrt} = \beta\, p_{hrt} + H + R + T + (H \times R) + (H \times T) + (R \times T) + \varphi_{hrt}, \tag{13}$$

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 4 reports the wage 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 any issues related to attenuation bias. Therefore, it is not be surprising that the coefficient $\hat{\beta}$ reported in

Table 4 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 wage effect using the Statistics Canada file is -0.053, with a standard error of 0.037. In the United States, the estimated wage 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 attenuated by measurement error. The corrected coefficients reported in the table confirm our suspicions. Row 8 of Table 4 shows that the simplest (and probably more reliable) 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. Even abstracting from the possibility that spatial correlations are biased downwards because of equilibrating adjustments, the substantial attenuation bias implicit in the estimated coefficients has led to a significant understatement of the observed wage impact of immigration across local labor markets.[20]

---

[20] Card's (1991) influential study of the Mariel flow is not susceptible to the type of measurement 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 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 a phantom immigrant influx ("The Mariel Boatlift That Didn't Happen") had a *significant and adverse* impact on labor market conditions in Miami. It is obvious, therefore, that confounding factors in Card's difference-in-differences analysis are not well understood and drive the results.

It is also worth noting that the back-of-the-envelope correction given by equation (7) leads to a larger adjustment than the correction that calculates the mean of the hypergeometric sampling error. Row 9 shows that using the method implied by equation (10) roughly doubles the estimated wage impact of immigration, both in the United States and Canada. The difference in the size of the adjustment implied by the two methods arises partly because of the relatively large number of cells that have a zero immigrant share and hence contribute nothing to the calculation of the mean sampling error.[21]

Not surprisingly, the bias in the estimated wage impact of immigration becomes substantially worse when we consider smaller samples. Consider, for instance, the PUMF sampling rate in Canada—the sampling rate that comprises the publicly available data. 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 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.

Note also that it is often not possible to use the various methods of correction to adjust the inconsistent coefficient $\hat{\beta}$ because the cell size used to calculate the immigrant share is so small—even at the 5/100 sampling rate. For instance, we find that because the predicted sampling error exceeds the adjusted variance of the immigrant share for many of the replications, the average corrected coefficient is wrong-signed or extremely negative, and has a very large

[21] The divergence between the two sets of corrections would be narrowed if we replaced the estimate of the immigrant share in cells that have a value of near-zero or zero with a value of 0.02 or 0.03.

standard deviation. There are limits, therefore, as to what can be learned from data that contains substantial measurement error.

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. In fact, the comparison of the average coefficient estimated in the 1/100 replications (-0.022) with the "true" impact predicted by the back-of-the-envelope correction in the 5/100 analysis (-0.170) suggests that the bias is on the order of 90 percent. Interestingly, the use of the immigrant share calculated in the 5/100 samples to estimate the coefficient β^* in the 1/100 replications does not fully recover the true effect. The estimated coefficient is -0.032, with an average standard error of 0.082.

Note also that the back-of-the-envelope correction leads to a wrong-signed corrected coefficient in the 1/100 replication because the predicted sampling error was often larger than the adjusted variance. We were also unable to implement the regression-based correction method to adjust the estimated wage impact of immigration at the metropolitan area level in the United States. The problem is that a regression estimated across local labor markets contains around 12,000 fixed effects. In the typical fixed-effects regression, the estimation problem implied by such large numbers of dummy variables is bypassed by simply differencing the data and

estimating the regression on the differenced data. The differencing "trick," however, cannot be applied to the errors-in-variables estimator. As a result, the regression-based method does not provide a practical way of correcting for measurement error when there are so many cells.

C. Instrumental Variables

Income-maximizing immigrants may cluster in particular (geographic or skill-based) labor markets because those are the markets that offer particularly high returns to the mobility costs incurred by the migrants. The immigrant share coefficient from an OLS wage regression 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; Ottaviano and Peri, 2005). 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.[22] In theory, these IV regressions could provide an alternative method for correcting for measurement 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—after all, 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

---

[22] 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, give 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.

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 IV regressions, leading both to weak instruments and to the violation of a key assumption in the classical measurement error model.

We document the sensitivity of the instrument to sampling error by estimating the generic first-stage regression:

$$p_{tk} = \delta\, p_{t-1,k} + \sum_h \gamma_h X_{tkh} + \varepsilon_{tk}\,, \tag{14}$$

where p_{tk} is the observed immigrant share for cell k in the current period and $p_{t\text{-}1,k}$ is the lagged share. As before, the observed immigrant shares are defined by: $p_{tk} = \pi_{tk} + u_{tk}$ and $p_{t\text{-}1,k} = \pi_{t\text{-}1,k} + u_{t\text{-}1,k}$, 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.[23]

Table 5 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. Consider the results obtained in the Canadian labor market. The coefficient of the lagged immigrant share in the large Statistics Canada file is 0.258, with a standard error of 0.085,

---

[23] 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.

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 that use 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 metropolitan area analysis for the United States, the coefficient $\hat{\delta}$ is already negative even at the 5/100 sampling rate.[24] 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 show that this meltdown occurs because there are measurement errors on both sides of the first-stage regression equation *and* these errors are correlated. Table 5 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

[24] 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 the 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).

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 samples, the first-stage regression model actually builds in a strong negative correlation in the errors between the two sides of the equation. 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:

$$(15) \qquad p_{t,k} - \overline{p}_{t,k} = \delta\,(p_{t-1,k} - \overline{p}_{t-1,k}) + \text{fixed effects} + \varepsilon,$$

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 readily apparent by considering the special case where the data consists of two cross-sections. We can then rewrite equation (15) as:

$$(15') \qquad p_{tk} - p_{t\text{-}1,k} = \delta\,(p_{t\text{-}1,k} - p_{t\text{-}2,k}) + \text{fixed effects} + \varepsilon.$$

The appearance of $p_{t-1,k}$ 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 (15′) 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 with a lagged dependent variable helps explain the pattern of estimated coefficients reported in Table 5. In particular, note that in the apparent absence of sampling errors (e.g., in the national-level regressions estimated either in the Canada Statistics or 5/100 U.S. Census files), the estimated coefficient $\hat{\delta}$ is strongly positive. Errors in the right-hand-side of equation (15′) 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 5 clearly indicate that the lagged immigrant share is not a valid instrument when the cell size is sufficiently small—even when we abstract from any conceptual issues. Put differently, IV regressions that use the lagged immigrant share as the instrument will not typically correct for the measurement error problems introduced by sampling error, particularly when the unit of analysis is a labor market defined by geography and skill.

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 immigrant-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 proposed and tested a new hypothesis that can account for the preponderance of weak estimated effects in the literature: the estimated wage impact of immigration is greatly 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 methodological exercise 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. Further, because the variable measured with error is a proportion, the properties of the hypergeometric distribution can be used to precisely characterize the nature of the attenuation bias in the estimated coefficients.

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 empirical 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 fraction of the workforce that is foreign-born can be 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 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 easily lead to a reappraisal of many regression-based stylized facts.

DATA APPENDIX: CONSTRUCTION OF CENSUS EXTRACTS AND VARIABLE DEFINITIONS

***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.

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"). We rely on years of grade school (*highgrad*), vocational training (*training*), and years of post-secondary education below university (*otheredu*) to make the 1971 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. 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.

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.

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. "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics* 1 (March 1973): 49-59.

Altonji, Joseph G. and Card, David. "The Effects of Immigration on the Labor Market Outcomes of Less-Skilled Natives," in John M. Abowd and Richard B. Freeman, eds., *Immigration, Trade, and the Labor Market*. Chicago: University of Chicago Press, 1991, pp. 201-234.

Angrist, Joshua D., and Alan B. Krueger, Empirical Strategies in Labor Economics, in Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3A, 1999.

Aydemir, Abdurrahman and George J. Borjas. "A Comparative Analysis of the Labor Market Impact of International Migration: Canada, Mexico, and the United States," *Journal of the European Economic Association*, forthcoming 2007.

Bloch, Farrell E. "Measurement Error and Statistical Significance of an Independent Variable," *American Statistician* 32(1) (February 1978): 26-27.

Bohn, Sarah and Seth Sanders. "Refining the Estimation of Immigration's Labor Market Effects," University of Maryland Working Paper, February 2005.

Bonin, Holger. "Is the Demand Curve Really Downward Sloping?" IZA Working Paper, May 12, 2005.

Borjas, George J. "Immigrants, Minorities, and Labor Market Competition," *Industrial and Labor Relations Review* 40 (April 1987): 382-392.

Borjas, George J. "The Labor Demand Curve *Is* Downward Sloping: Reexamining the Impact of Immigration on the Labor Market," *Quarterly Journal of Economics* 118 (November 2003): 1335-1374.

Borjas, George J. "Native Internal Migration and the Labor Market Impact of Immigration," *Journal of Human Resources*, forthcoming 2006.

Borjas, George J., Richard B. Freeman, and Lawrence F. Katz. "How Much Do Immigration and Trade Affect Labor Market Outcomes?" *Brookings Papers on Economic Activity* (1997): 1-67.

Bound, John, Charles C. Brown, and Nancy Mathiowetz. "Measurement Error in Survey Data," in *Handbook of Econometrics*, edited by Edward E. Learner and James J. Heckman. New York: North Holland Publishing, 2001, pp. 3705-3843.

Card, David. "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43 (January 1990): 245-257.

Card, David. "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration," *Journal of Labor Economics* (January 2001): 22-64.

Filer, Randall K. "The Impact of Immigrant Arrivals on Migratory Patterns of Native Workers," in George J. Borjas and Richard B. Freeman, eds., *Immigration and the Work Force: Economic Consequences for the United States and Source Areas*. Chicago: University of Chicago Press, 1992, pp. 245-269.

Freeman, Richard B. "Longitudinal Analyses of the Effect of Trade Unions," *Journal of Labor Economics* 2(1) (January 1984): 1-26.

Frey, William. "Immigration Impacts on Internal Migration of the Poor: 1990 Census Evidence for U.S. States," *International Journal of Population Geography* 1 (1995): 51-67.

Friedberg, Rachel M. and Jennifer Hunt. "The Impact of Immigration on Host Country Wages, Employment and Growth," *Journal of Economic Perspectives* 9 (Spring 1995): 23-44.

Garber, Steven and Steven Keppler, "Extending the Classical Normal Errors-in-Variables Model," *Econometrica* 48(6) (September 1980): 1541-1546.

Grossman, Jean Baldwin. "The Substitutability of Natives and Immigrants in Production," *Review of Economics and Statistics* 54 (November 1982): 596-603.

Gurak, Douglas T. and Mary M. Kritz. "The Interstate Migration of U.S. Immigrants: Individual and Contextual Determinants." *Social Forces* 78(3) (March 2000):1017-39.

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

Kane, Thomas J., Cecilia E. Rouse, and Douglas Staiger. "Estimating Returns to Schooling When Schooling is Misreported," NBER Working Paper No. 7235, July 1999.

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

LaLonde, Robert J. and Robert H. Topel. "Labor Market Adjustments to Increased Immigration," in John M. Abowd and Richard B. Freeman, eds., *Immigration, Trade, and the Labor Market*. Chicago: University of Chicago Press, 1991, pp. 167-199.

Levi, Maurice D. "Errors in the Variables Bias in the Presence of Correctly Measured Variables," *Econometrica* 41(5) (September 1973): 985-986.

Lewis, Ethan. "Immigration, Skill Mix, and the Choice of Technique," Federal Reserve Bank of Philadelphia Working Paper #05-08, May 2005

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

Ottaviano, Gianmarco I.P. and Giovanni Peri. "Rethinking the Gains from Immigration: Theory and Evidence from the U.S." NBER Working Paper No. 11672, September 2005.

Pischke, Jörn-Steffen and Johannes Velling. "Employment Effects of Immigration to Germany: An Analysis Based on Local Labor Markets," *Review of Economics and Statistics* 79 (November 1997): 594-604.

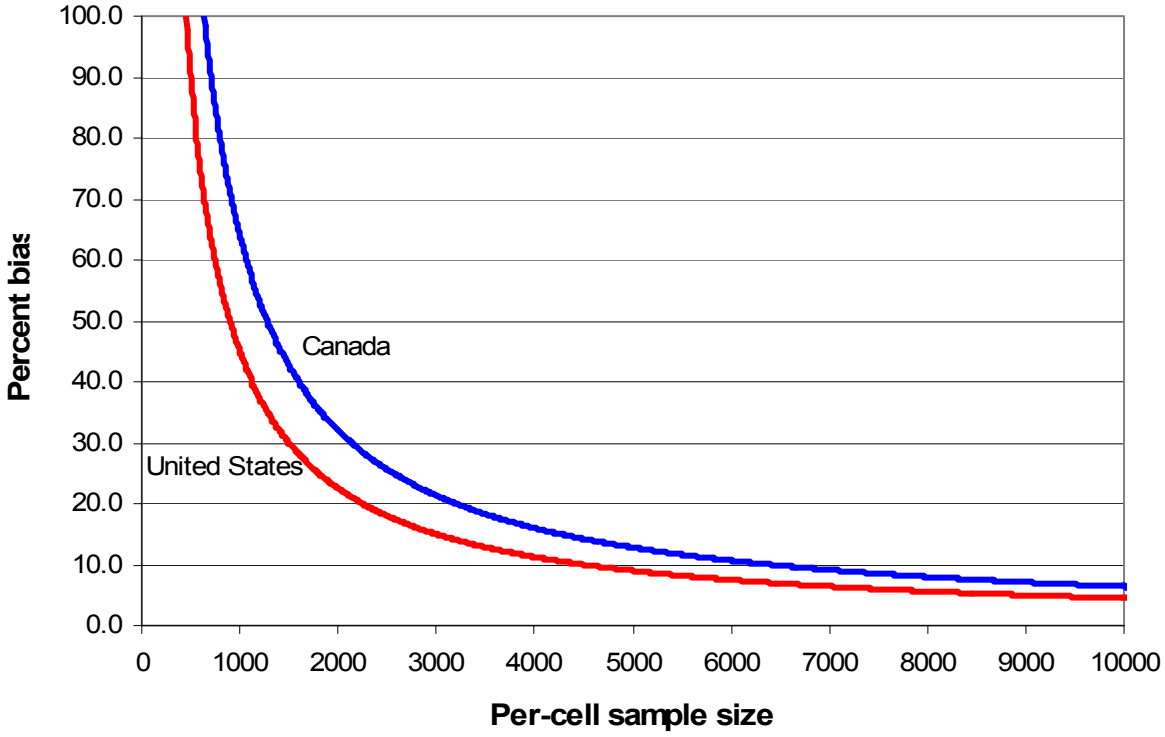
Saiz, Albert. "Room in the Kitchen for the Melting Pot: Immigration and Rental Prices," *Review of Economics and Statistics* 85 (August 2003): 502-521.

Schoeni, Robert F. "The Effect of Immigrants on the Employment and Wages of Native Workers: Evidence from the 1970s and 1980s," unpublished paper, The RAND Corporation, March 1997.

Smith, James P. and Barry Edmonston, editors. *The New Americans: Economic, Demographic, and Fiscal Effects of Immigration*. Washington, D.C.: National Academy Press, 1997.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics* 20 (October 2002): 518-529.

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



Notes: The simulation for Canada assumes that the mean immigrant share is 0.2; the variance of the immigrant share across national-level local 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

Notes: 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

	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. R^2 of auxiliary regression	0.967	0.965	0.960	0.953	0.845	0.590
Using "large sample" share						
5. β^*	---	-0.505	-0.501	-0.499	-0.466	-0.384
6. Standard error of β^*	---	0.209	0.226	0.241	0.485	1.475
7. Standard deviation of β^*	---	0.049	0.093	0.126	0.405	1.353
Corrected coefficients:						
8. Using mean immigrant share	-0.520	-0.531	-0.524	-0.638	1.174	0.044
9. Standard deviation of row 8	---	0.064	0.132	0.241	15.647	1.652
10. Using mean of binomial error	-0.521	-0.538	-0.590	-0.689	0.192	-0.400
11. Standard deviation of row 10	---	0.065	0.151	0.265	7.693	14.139
12. Regression method	-0.520	-0.531	-0.524	-0.638	n.a.	n.a.
13. Std. error of regression coeff.	0.118	0.138	0.167	0.259	n.a.	n.a.
United States:						
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. R^2 of auxiliary regression	---	0.974	---	0.973	0.964	0.883
Using "large sample" share						
5. β^*	---	---	---	-0.488	-0.497	-0.498
6. Standard error of β^*	---	---	---	0.228	0.291	0.631
7. Standard deviation of β^*	---	---	---	0.045	0.171	0.534
Corrected coefficients:						
8. Using mean immigrant share	---	-0.496	---	-0.506	-0.642	5.794
9. Standard deviation of row 8	---		---	0.060	0.320	89.464
10. Using mean of binomial error	---	-0.496	---	-0.507	-0.658	-0.287
11. Standard deviation of row 10	---	---	---	0.060	0.331	9.328
12. Regression method	---	-0.497	---	-0.506	-0.646	n.a.
13. Std. error of regression coeff.	---	0.179	---	0.189	0.372	n.a.

Notes: 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 Canadian labor market has 240 cells; the analysis of the U.S. labor market has 200 cells.

Table 3. 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

Notes: 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 4. 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. Using mean immigrant share	-0.112	0.328	0.065	0.009
9. Standard deviation of row 8	---	0.921	0.196	0.099
10. Using mean of binomial error	-0.090	-0.131	-0.168	-.737
11. Standard deviation of row 10	---	0.192	1.746	23.320
12. Regression method	-0.109	n.a.	n.a.	n.a.
13. Std. error of regression coefficient	0.061	n.a.	n.a.	n.a.
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. Using mean immigrant share	---	-0.170	---	0.036
9. Standard deviation of row 8		---	---	0.031
10. Using mean of binomial error	---	-0.084	---	-0.132
11. Standard deviation of row 10		---	---	0.117
12. Regression method	---	n.a.	---	n.a.
13. Std. error of regression coefficient	---	n.a.	---	n.a.

Notes: 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 5. 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	---	---

Notes: 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; $\hat{\delta}(p_t, \pi^*_{t-1})$ 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.