Canadian Studies in Population 40, no. 3–4 (2013): 166–173.

Studying immigrant earnings with alternative data sources and model specifications: A reply to DeVoretz's comment

Feng Hou*

Professor DeVoretz raises several important issues that many empirical researchers have to critically consider when facing choices of alternative data sources, measures, and model specifications. Such choices are often difficult to make, as each alternative has advantages and disadvantages. When a perfect choice is not available, one would want to have a sense of how much it would matter to have a better option. In responding to the reviewer's concerns, I will attempt to provide such a sense on issues related to my paper. I will argue that that the limitations associated with my choices would not substantially affect the answers to my research question.

Census vs. other data sources

Professor DeVoretz suggests that the lack of immigrant entry class information in the Census limits what one can learn about the earnings gap between immigrants and the Canadian-born. While how immigrants were admitted into Canada is not known from the Census, Census micro-data files remain a valuable source to study the economic integration of immigrants. The large sample size, comprehensive socio-demographic variables, and consistent measures of multi-year census files make it possible to study long-term trends in immigrant earnings and some determinants affecting these trends. Other data sources, such as the Longitudinal Survey of Immigrants to Canada (LSIC) and the Longitudinal Immigration Database (IMDB), do have information on immigrant entry class and other immigrant characteristics prior to or at entry, but they have their own disadvantages. For instance, the LSIC follows immigrants for only four years, based on the experience of the 2000–2001 landing cohort. As my paper shows, immigrants who landed in this period were highly concentrated in IT professions and bore the brunt of the IT bust of the early 2000s. Their experience is hardly representative of earlier and later arrival cohorts. The IMDB, or the combined Longitudinal Administrative Databank (LAD) and IMDB, are rich on immigrant characteristics at entry. The main drawback, however, is that critical variables, including education and race/ethnicity, are not available for the Canadian-born population. Consequently, it is not possible to examine earnings gaps between observationally equivalent immigrants and the Canadian-born. Furthermore, there is no information on working time and occupation in the IMDB; therefore, one cannot examine weekly earnings and earnings gaps by occupation on the basis of the IMDB, while it was possible to do so, as I did, with Census data.

How much does it matter that the Census does not contain immigrant class information? Immigration class is certainly a major determinant of immigrant earnings. Not only do immigrants in different entry classes (e.g., family class vs. skilled workers) differ in their characteristics observed in

* Feng Hou, Social Analysis Division, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa ON K1A 0T6. Email: feng.hou@statcan.gc.ca.

the Census—primarily in terms of education, official language, and age—but they may also differ in unobserved abilities, motivations, and preparedness for the Canadian labour market. However, it remains an empirical question whether not controlling for immigrant class would bias the estimated changes in entry earnings across cohorts. This question cannot be answered with Census data. Using the IMDB to examine changes in earnings across cohorts, with and without immigrant class, will shed light on this issue. Note that, with the IMDB, one can examine only changes in earnings among immigrants: earnings gaps between immigrants and observationally equivalent Canadian-born cannot be observed with the IMDB.

Table 1. Regression models predicting log annual earnings of immigrant men, age 25–64.

	Model 1	Model 2
	Coefficient	Coefficient
Cohort 1981–1985	0.135***	0.125***
Cohort 1986–1990	0.177***	0.185***
Cohort 1991–1995	−0.120***	−0.089***
Cohort 2001–2005	−0.098***	−0.098***
Years since immigration	0.446***	0.454***
Years since immigration squared	−0.005***	−0.005***
Cohort 1981–1985* years since immigration	−0.027***	−0.026***
Cohort 1986–1990* years since immigration	−0.065***	−0.065***
Cohort 1991–1995* years since immigration	−0.006***	−0.005***
Cohort 2001–2005* years since immigration	0.021***	0.021***
Years of foreign experience	−0.004***	−0.004***
Less than high school	−0.228***	−0.121***
High school graduation	−0.164***	−0.091***
Some post-secondary	−0.149***	−0.082***
Graduate degrees	0.062***	0.037***
French	−0.165***	−0.182***
Both English and French	−0.014***	−0.049***
Neither English nor French	−0.143***	−0.094***
Northern and Western Europe	0.075***	0.004
Southern and Eastern Europe	−0.374***	−0.404**
Africa	−0.521***	−0.519***
East Asia	−0.677***	−0.746***
South Asia	−0.605***	−0.630***
Southeast Asia	−0.490***	−0.490***
West Asia/Middle East	−0.708***	−0.725***
Caribbean, Central and South America	−0.484***	−0.488***
Other countries	−0.230***	−0.242***
Family class		−0.265***
Business class		−0.336***
Spouses and dependants of skilled workers		−0.196***
Refugees		−0.319***
Other class		−0.270***
Marital status	included	included
Geographic location fixed effects	included	included
Sample size	4,224,170	4,224,170
R-squared	0.219	0.237

Source: Longitudinal Immigrant Database

Note: * significant at p <0.05, ** p<0.01, *** p< 0.001. The reference category is the 1996–2000 cohort for the cohort variables, Bachelor's degree for education, English for mother tongue, the US for source regions, and Skilled worker principal applicants for immigrant class.

Table 1 presents model estimates based on the IMDB. The sample includes immigrant men who landed from 1981 to 2005 at their prime work age (25 to 54). The earnings are observed in the years from 1982 to 2009; the latest arrivals in the selected sample (those arrived in 2005) therefore have a minimum of four years of observations. To be consistent with my paper, only recent immigrants (those who had resided in Canada five years or less) are used in the models. The outcome is log annual paid employment earnings. To save space, the coefficients of marital status and geographic regions of residence are not presented in the table.

Three points can be summarized from the results in Table 1. First, there are large earnings differences among immigrant classes even when other important immigrant characteristics are taken into account. The addition of immigrant class increases the model R-squared from Model 1 to Model 2 by 2 percentage points. Second, the addition of immigrant class in Model 2 substantially changes some coefficients of education, mother tongue, and source region. This finding reflects that immigrant class is correlated with these three variables. As a result, when immigrant class is not controlled for, education, mother tongue, and source region capture some of the effect of immigrant class on earnings. Third, the addition of immigrant class in Model 2 does not alter the trends in earnings across cohorts. The coefficients of several cohort dummy variables change somewhat in magnitude from Model 1 to Model 2, particularly the coefficient associated the 1991–1995 cohort. The overall trend, however, remains the same. The results on the overall trends are directly relevant to the focus of my paper.

My analysis in this paper does not rely on source country as the primary control for immigrant characteristics, although the inclusion of this factor in the analysis is useful in controlling for differences in unobserved characteristics among immigrants from different countries (e.g., university-educated immigrants from China earned less than university-educated immigrants from the UK). In addition to source region, I also control for changes in education, potential years of Canadian and foreign work experience, marital status, full-time/part-time status, visible-minority status, location of residence, mother tongue, and self-reported official language.

Earnings measures

Dr. DeVoretz considers hourly wages to be a better measure than monthly or yearly earnings of the disadvantage (or discrimination, in the phrase of the reviewer) experienced by recent immigrants in the labour market. Because hourly earnings are not available from the Census data, I used weekly earnings and controlled for full-time status. The Census collects information on total wages and salary, and weeks worked in the calendar year prior to the census date. From these two pieces of information, one can derive weekly earnings which eliminate variations among workers in weeks worked. The Census also asks whether the typical weeks worked are full-time (at least 30 hours a week). Controlling for full-time status to a large extent reduces variations among workers in hours worked per a typical week.

How much does it matter whether hourly wages instead of weekly wages are used as the outcome measure? While hourly wages cannot be derived from either the Census or the IMDB, the Labour Force Survey (LFS) does contain both weekly and hourly wages. The LFS is Canada's official source of monthly estimates of total employment and unemployment; it collects data on immigration status and country of study starting from 2006. By pooling data from 2006 to 2012, one can have a large sample of recent immigrants. Given the short span of years with relevant variables, the data set is not yet ideal for examining changes in wages across immigrant cohorts. Consequently, I focus on the cross-sectional wage gap between recent immigrants (defined here as those arrived since 2006) and the Canadian-born to demonstrate the difference resulting from using hourly wages vs. weekly wages.

Table 2. Regression models predicting hourly and weekly wage gaps between recent immigrant and the Canadian-born paid workers, age 25–64.

	Log hourly wages	Log weekly wages	Log weekly wages
	Coefficient	Coefficient	Coefficient
Recent immigrants	-0.357***	-0.361***	-0.393***
Full-time status	Not included	Included	Not included
Sample size	187,830	187,830	187,830
R-squared	0.198	0.364	0.149

Source: Labour force survey 2006–2012, May and November data.
Note: * significant at p <0.05, ** p<0.01, *** p< 0.001. All models also include education, age, tenure, and geographic regions of residence.

The results in Table 2 show that using hourly wages or weekly wages as the outcome makes little difference in the wage gap between recent-immigrant and Canadian-born male paid workers, as long as full-time status is controlled for in the model for weekly wages. The wage gap is about 3 percentage points larger when weekly wages are used as the outcome without controlling for full-time status. Census data also show that the earnings gap between recent immigrants and the Canadian-born is larger when annual earnings are used without controlling for weeks worked than when weekly earnings are used as the outcome.

The above results suggest that recent immigrants are disadvantaged not only in wage rates but also in working time. When the hourly wage rate is used as the outcome, or differences in working time are fully taken into account, the earnings gap experienced by recent immigrants manifests primarily through differential earnings structures within jobs. The corresponding assumption is that working time is not differentially constrained across groups, but rather is a voluntary choice. By contrast, when differences in working time are not controlled for, any earnings disadvantage of recent immigrants would reflect the combined effect of differential allocation of working time and earnings structures within jobs. If sorting into part-time jobs or fewer work hours is involuntary, focusing on hourly wages would underestimate the disadvantages experienced by recent immigrants in the labour market. In terms of economic well-being, higher annual earnings matter more than a higher price for labour.

Educational quality and location of study

As noted by Professor DeVoretz, without an actual measure, any discussion of educational quality is dubious. I agree. To my knowledge, educational quality is not available in any large survey in Canada (with the exception of some literacy and numeracy scales, which do not reflect general educational quality). Nevertheless, it is still possible to take into account how the labour market rewards the same level of education differently for immigrants and the Canadian-born, and for immigrants from various regions. In my analysis, I included an interaction term between immigrant status and education. This allows different earnings returns to the same level of education for immigrants and the Canadian-born. The inclusion of detailed source countries is another indirect way to capture the difference in educational quality among immigrants from different countries.

More importantly, the cited literature (Aydemir and Skuterud 2005; Green and Worswick 2010) suggests that the long-term decline in immigrants' relative earnings had little to do with immigrants' educational quality, even though at a given point in time immigrants from countries with higher educational quality perform better than those from countries with lower educational quality. The purpose of my paper is to account for changes in immigrants' entry earnings, not to predict immigrant earn-

ings at a fixed time point. The latter type of analysis can be found in a report discussing the effect of country of study on immigrant earnings (Frenette et al. 2008).

The reviewer seems to suggest that location of study is equivalent to education quality. The two are certainly related, but the strength of this relationship is open to debate. The benefits of receiving university degrees from Canada or other developed countries may not be confined to higher educational quality; it may also include improved knowledge about labour market opportunities in modern economies, improved ability in the host-country's official languages, and social networks. There is also an issue of selectivity. Immigrants who were born in less developed countries but received their education in Canada or in other developed countries may differ from other immigrants in many ways. These possible mechanisms through which location of study affects earnings are difficult to disentangle. In large representative national data sources, such as the Census and the LFS, information is available only on the country of study—the country where immigrants received their highest level of education.

Empirically, how much would information on country of study improve our understanding of immigrant earnings? Again, using the pooled 2006–2012 LFS data and including immigrant men who had been in Canada for 1 to 20 years, I have estimated regression models to predict log hourly wages among immigrant male paid workers. Country of study is coded as five groups: Canada, the US, UK/Australia/New Zealand, Continental Europe, and Other Countries. When the model includes education, years since immigration and its squared term, age at immigration, as well as job tenure and province of residence, the model R-squared is 0.216, implying that about 22 per cent of the variation in hourly wages among immigrant workers is accounted for by the included variables. When region of study is added to the model, the model R-squared increased by 0.013. Thus, beyond the explanatory power of other demographic factors, the region of study explains an extra 1.3 per cent of the variation in immigrant earnings.

Entry earnings and assimilation effects

Professor DeVoretz argues that my focus on entry earnings fails to incorporate any catch-up trends over time. My approach is certainly divergent from the conventional approach of estimating the cohort effect and assimilation effect in a single model. Let me first explain why it is problematic, with the common approach of pooling cross-sectional data over a long time span, to estimate immigrant cohort and assimilation effects in a single regression model. When a synthetic cohort is "followed" over a long time, its size and composition may change, as a result of either the addition of childhood immigrants who were not included when entry earnings were estimated or attrition in later periods resulting from withdrawal of older immigrants from the labour force, emigration, or death. Consequently, both the cohort effect (cohort differences in entry earnings) and assimilation effect (earnings growth with years since immigration) are likely estimated with bias.

The problem with the addition of younger immigrants in the later censuses can be mitigated by restricting the range of age at immigration (for example, to 25–40) in order to ensure that the same age cohort of immigrants is followed in the multi-year cross-sectional data. However, it remains impossible to overcome the attrition issue with multi-year cross-sectional data. A Canadian study suggests that about one-third of working-age immigrant men moved out of Canada within 20 years after immigrating to Canada (Aydemir and Robinson 2006). The attrition issue can be addressed only with longitudinal data such as the IMDB (e.g., Picot and Patrizio 2012).

There is a more serious problem, which is little known to the immigration research community. It is related to the severe bias that occurs when model estimates are based on extrapolation of limited data points. The conventional approach, which estimates cohort and assimilation effects simultan-

eously, pools successive arrival cohorts with different lengths of observation, ranging from a couple of years to as long as two decades. With such data, a model would include a series of dummies for cohorts, years since immigration, and the interaction terms between cohorts and years since immigration. The size and significance of the coefficients of these interaction terms would be used to infer whether cohorts with lower entry earnings would catch up with cohorts with higher entry earnings. A critical issue concerning this approach is that the earnings growth profiles of the cohorts with just a few years of observation are estimated from the extrapolation of observed short-term trends, even though these trends may not continue. Imposing long-term growth profiles on cohorts with short periods of observation would result in biased estimates of entry earnings and growth profiles for all cohorts in the model.

An empirical example can illustrate how serious the problem could become. With the IMDB, which contains earnings data from 1982 to 2009, I estimate two earnings models for three five-year arrival cohorts of immigrants: 1981–1985, 1986–1990, and 1991–1995. The first model uses earnings data up to 1996, pretending the observation after 1996 is not available. This is like the conventional approach, by which the last cohort has 1 to 5 years of observation, and thus its earnings growth is extrapolated in the estimation. The second model uses earnings data up to 2009, so that each cohort has at least 14 years of actual observation. Table 3 presents the model estimates, and Figure 1 and Figure 2 plot the entry earnings and the growth profiles predicted from these models.

Table 3. Regression models predicting log annual earnings of immigrant men with different lengths of observation.

	Earnings observed from 1982 to 1996	Earnings observed from 1982-2009
	Model 1	Model 2
	Coefficient	Coefficient
Cohort 1981–1985	0.325***	0.152***
Cohort 1986–1990	0.295***	0.096***
Years since immigration	0.206***	0.093***
Years since immigration squared	-0.007***	-0.002***
Cohort 1981-1985*years since immigration	-0.060***	-0.007***
Cohort 1986-1990*years since immigration	-0.073***	-0.007***
Sample size	2,982,100	6,905,420
R-squared	0.259	0.255

Source: Longitudinal Immigrant Database.

Note: * significant at p <0.05, ** p<0.01, *** p< 0.001. The reference group for cohort dummy variables is the 1991–1995 cohort. All models also include foreign experience, education, mother tongue, source regions, immigrant class, marital status and geographic regions of residence as specified in Model 2 of Table 1.

Comparing Model 1 with Model 2, and Figure 1 with Figure 2, it becomes clear that the conventional approach with incomplete observation (as in Model 1 and Fig. 1) generates three types of bias. First, it distorts the earnings growth profiles for all arrival cohorts. The growth profiles based on an incomplete observation period have a steep bowl shape, while the profiles based on at least 14 years of observation are close to linear. Put differently, the estimates based on incomplete data amplify both the acceleration of earnings growth in the initial years after immigration and the later deceleration of earnings growth.

Second, the estimates based on an incomplete observation period greatly exaggerate the earnings growth of the 1991–1995 cohort. The estimates imply that this cohort would catch up with the two previous cohorts within five to six years after immigration. This cohort arrived during a recession and had

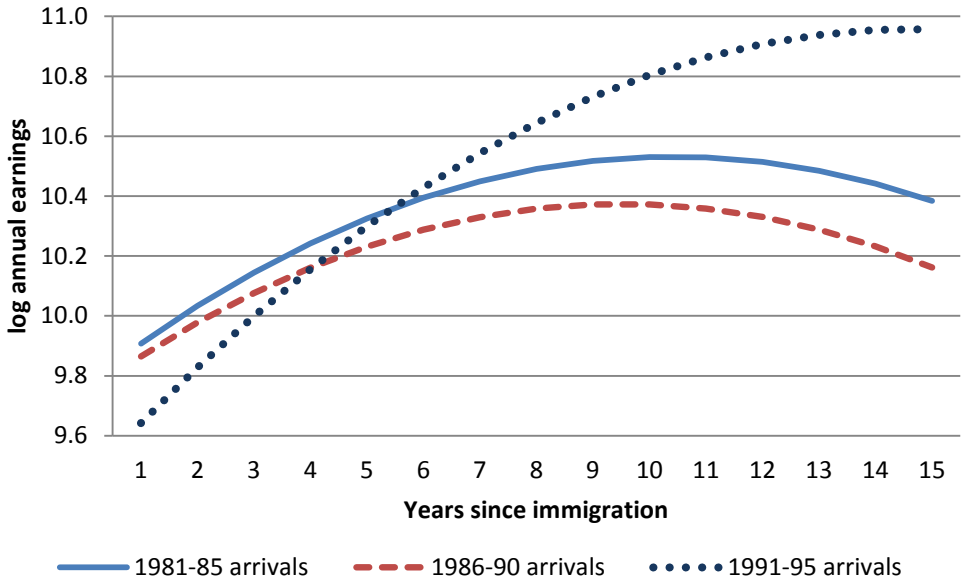

Figure 1. Predicted earnings profiles of immigrant men based on observation covering the 1982–1996 period.

low entry earnings, but achieved fast earnings growth in the following period of economic expansion. Extrapolation of this brief period of observation would inflate this cohort's long-term growth pattern. The estimates based on observation up to 2009 show that the 1991–1995 cohort had not caught up with the 1981–1985 cohort by the fifteenth year after immigration. Conversely, the estimates based on incomplete data underplay the growth rate of the 1986–1990 cohort. This cohort came during a period of economic expansion, followed by a period of recession and slow recovery. Extrapolation of this short-term trend deflates this cohort's long-term growth pattern. When estimates are based on a long period of observation, the earnings profiles of the three cohorts are almost parallel (Fig. 2).

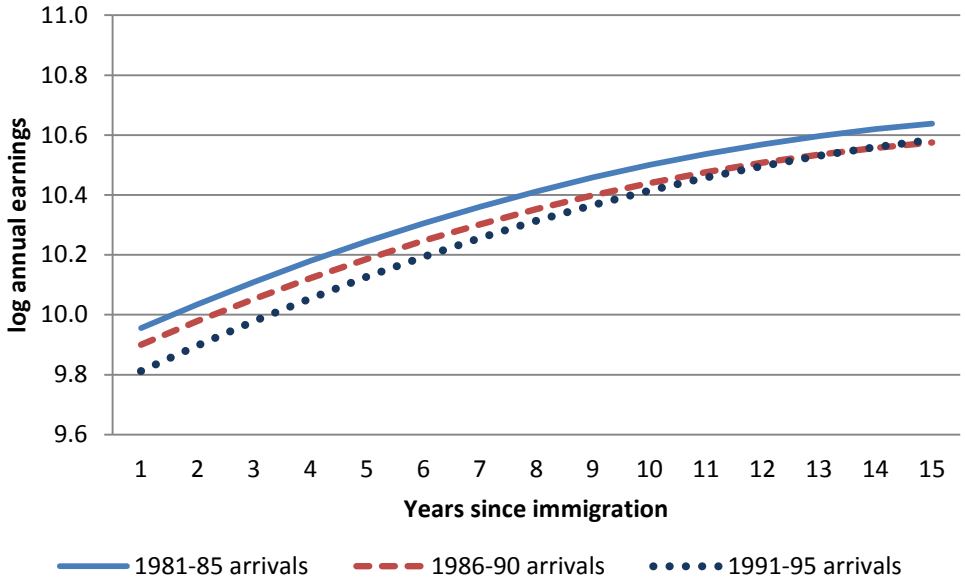

Figure 2. Predicted earnings profiles of immigrant men based on observation covering the 1982–2009 period.

Third, the estimates based on incomplete observation greatly overstate the entry earnings gap across cohorts, by as much as 100 per cent. According to the estimates from an incomplete period of observation, the 1981–1985 cohort would have entry earnings 32.5 per cent higher than those of the 1991–1995 cohort. With a long period of observation, the gap reduces to 15.2 per cent.

The lesson learned from this example is that sacrifices have to be made in considering cohort and assimilation effects. If the research focus is on assimilation effects, only cohorts with a long enough observation period should be included. In that case, there is not much one can say about more recent arrivals. If the focus is on cohort effects, particularly when one examines how recent cohorts perform differently from earlier cohorts, one should only compare their entry earnings without imposing any long-term earnings growth profiles. The latter is the choice that I made for my paper.

Not only is focusing on entry earnings necessary for the purpose of my study, this approach has other advantages, too. As discussed in the paper, it allows for a very flexible model that includes changes in the returns to both Canadian work experience and education as explanatory factors. It also allows the application of a decomposition technique for evaluating the relative contribution of each explanatory variable to changes in earnings gaps across cohorts.

In sum, using census data to answer the research questions of this study is quite reasonable. The reviewer is correct in stating that the Census has limitations. His comments compelled me to gauge the extent to which these limitations matter to specific research issues. I have concluded that these limitations do not negate the overall findings of my paper. I thank Dr. DeVoretz and the editor, Dr. Trovato, for giving me the opportunity to do so.

References

Frenette, Marc, Feng Hou, Rene Morissette, Ted Wannell, and Maryanne Webber. 2008. *Earnings and Incomes of Canadians Over the Past Quarter Century: 2006 Census*. Ottawa: Statistics Canada. Catalogue No. 97-563-X.

Picot, G., and P. Piraino. 2012. *Immigrant Earnings Growth: Selection Bias or Real Progress?* Analytical Studies Branch Research Paper Series No. 340. Ottawa: Statistics Canada. Catalogue no. 11F0019M.