I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14217

Immigrants' Economic Performance and Selective Outmigration: Diverging Predictions from Survey and Administrative Data

Charles Bellemare
Natalia Kyui
Guy Lacroix

MARCH 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14217

Immigrants' Economic Performance and Selective Outmigration: Diverging Predictions from Survey and Administrative Data

Charles Bellemare
Laval University and IZA

Natalia Kyui
Bank of Canada

Guy Lacroix
Laval University and IZA

MARCH 2021

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Immigrants' Economic Performance and Selective Outmigration: Diverging Predictions from Survey and Administrative Data*

We show that survey and administrative data-based estimates of a panel data model of earnings, employment, and outmigration yield very different qualitative and quantitative predictions. Survey-based estimates substantially overpredict outmigration, in particular for lower performing immigrants. Consequently, employment and earnings of immigrants who remain in the country are overpredicted relative to model predictions from administrative data. Importantly, estimates from both data sources find opposite self-selection mechanisms into outmigration. Differences hold despite using the same cohort, survey period, and observable characteristics. Differences in predictions are driven by difficulties of properly separating non-random sample attrition from selective outmigration in survey data.

JEL Classification: C33, J31, J15, J61

Keywords: sample attrition, outmigration, measurement errors, employment and earnings

Corresponding author:
Guy Lacroix
Department of Economics
Université Laval 1025, avenue des Sciences-Humaines
Québec (Québec)
Canada G1V 0A6

E-mail: guy.lacroix@ecn.ulaval.ca

* The authors thank Statistic Canada for data access. The views expressed in this paper are those of the authors and not necessarily those of the Bank of Canada.

1 Introduction

The importance of immigration in many societies has prompted significant academic research measuring the economic performance of immigrants, notably focusing on their employment levels, earnings, and duration of stay in the host country (see Dustmann and Görlach (2015) for a recent overview of the literature).[1] In Canada, more than 25% of the labour force are foreign-born.[2] Immigration has been the main driver of the population growth in recent years, contributing around 70% to its annual increase, and is expected to account for a larger proportion of population and labour force growth in the near future.[3] Among the population with a university degree, the share of immigrants is even larger and has been increasing yet even faster.[4] The case of Canada is by no means an exception, and many other OECD countries have experienced similar trends over the last decades (see Ferrie and Hatton (2015)).

Modern migration patterns suggest that a large fraction of the foreign-born population will at some point return to their home country or move to another host country. Indeed, according to OECD data the ratio of outflows to inflows over a 10-year period varies between 9.8% in Australia to as much as 42.6% in the UK and 51.5% in Switzerland (Dustmann and Glitz, 2011, Table 4.8). This is consistent with the estimates of Bijwaard (2010) who also finds that between 20–50% of immigrants to Western Europe will eventually return home. Thus, re-emigration is by no means a marginal phenomenon. Data for Canada suggest that as many as 35% of young, male, working-age immigrants will have left permanently within 20 years after landing, the majority of which having done so during the first year (Aydemir and Robinson, 2008).[5] Factors that explain return and repeat migration, and hence duration of stay, include migrant characteristics, networks, migration costs, and immigration policy. More subjective factors, such as presence of family members, climate, culture, and lifestyle, also play a role (Dustmann and Weiss, 2007). Lastly, unemployment spells have been found to increase return probabilities, while reemployment spells typically delay outmigration (Bijwaard, Schluter and Wahba, 2014).

[1] A related literature has focused on the impact of immigration on wages and employment of native workers. See Edo, Ragot, Rapoport, Sardoschau, Steinmayr and Sweetman (2020) for a recent survey.

[2] Landed immigrants, aged 15 years and over, accounted for 25.7% of the labor force and for 26.2% of the working-age population in 2019. Source: Statistics Canada, Labour Force Survey, Table 14-10-0083-01.

[3] See Agopsowicz, Gueye, Kyui, Park, Salameh and Tomlin (2017) and Kustec (2012) for more details.

[4] In 2019, university degree holders accounted for 38.5% among landed immigrants and for 22.2% in the Canadian born population, aged 15 years and over. Source: Statistics Canada, Labour Force Survey, Table 14-10-0087-01.

[5] Using census and administrative data, Aydemir and Robinson (2008) find very similar retention rates for Canada. According to Damas de Matos and Parent (2019), as many as 18% (13%) of highly-skilled male (female) immigrants to Canada aged 30-39 will move to the United States within five years after landing.

To the extent that outmigration is selective and sizeable, it must be accounted for when investigating the economic performances of immigrants. A variety of approaches have been used to that end, namely non-structural (e.g., Borjas (1985)) and structural approaches (e.g., Bellemare (2007), Kirdar (2012)). Regardless of the approach followed, quality data on outmigration appears essential to determine whether successful or unsuccessful immigrants stay in their host country after landing. Quality of outmigration data will vary according to the data source used for the analysis. Administrative records such as those analyzed in this paper allow to track outmigration closely. On the other hand, survey data contain possibly more noisy measures of outmigration. To our knowledge, there has been no systematic analyses of how quality data on outmigration impacts measurements of the economic performance of immigrants in their host country.

In this paper we show that reliable information on outmigration is essential to properly measure the economic performance of immigrants in their host country. Our analysis exploits two Canadian databases that each document the presence of immigrants in the host country differently. The first is the Longitudinal Survey of Immigrants to Canada (LSIC), a comprehensive longitudinal survey administered by Statistics Canada which follows a sample of immigrants who landed in 2000-2001. The LSIC contains labour market information spanning three biennial waves (2001, 2003, 2005) as well as socio-economic characteristics and background information.[6] LSIC does not contain good quality indicators of outmigration, with the latter subsumed in sample attrition. The second is the Longitudinal Immigration Database (IMDB). The IMDB contains administrative data from Immigration, Refugees and Citizenship Canada derived from the Field Operations Support System (FOSS), and which was merged with income-tax files from Canada Revenue Agency. The IMDB covers all immigrants who have landed since 1980 and who have filed at least one tax return since 1982. It also contains socio-economic and background information similar to that contained in the LSIC.[7] Aydemir and Robinson (2008) compare IMDB attrition rates with corresponding immigrant cohort specific Census data and conclude that the vast majority of IMDB attrition reflects genuine outmigration, with temporary outmigration captured by immigrants not filling in a given year and permanent outmigration captured by permanent attrition from the database.

[6] Many academic publications have used LSIC in recent years (Sweetman and Warman, 2013; Warman, Sweetman and Goldmann, 2015; Imai, Stacey and Warman, 2019).

[7] The IMDB has also been used extensively in recent years to investigate different issues concerning immigrants in Canada (Green and Worswick, 2010; Picot and Piraino, 2013; Warman, Worswick and Webb, 2016).

Our analysis restricts the IMDB administrative database so as to generate an isomorphic sample matching the LSIC survey design. This allows to follow the exact same immigrant cohort, over the same period, and using the same observable characteristics. Therefore, the key difference between the two is their ability to follow an immigrant over time. Indeed, we show that our samples from LSIC and IMDB are very similar in terms of observed characteristics in the first year. However, the data suggests that the attrition rates computed from the LSIC across the three survey waves are 3 to 6 times greater than those computed from the IMDB data over the same three waves. The IMDB is presumably a more reliable source of information to investigate the economic performance of immigrants living in Canada since it better tracks residency.[8] Yet, outmigration may still be misclassified when using administrative data. Our empirical strategy consists in accounting explicitly for potential measurement error in outmigration as in Bellemare (2007).[9] We model employment, earnings, and (mismeasured) outmigration jointly using each database separately.

Predicted outcomes from the two samples differ substantially. Namely, the survey-based estimates depict an overoptimistic picture for the labour market performance of immigrants remaining in Canada, largely because of a sizeable overestimation of the proportion of lower performing immigrants leaving the country. To illustrate, the survey-based estimates overpredict outmigration rates in the first five years upon landing threefold, and up to fivefold in some specific cases (province-age-education levels). Consequently, predicted shares of immigrants remaining in Canada for all periods are largely underestimated, and especially so for labour market trajectories involving periodic unemployment. Indeed, the predicted share of immigrants employed in all three waves is underestimated by 8% on average and as much as 23% for some simulated cases. The predicted proportions of immigrants employed only in one or in any two waves are likewise underestimated by 28% and 25%, respectively (and by as much as 56% and 43% in some cases). Further, the proportion of immigrants never employed in all three waves is underestimated on average by 38% using LSIC data (and by up to 68% for some cases). Moreover, the outmigration overprediction in survey-based estimates affects to a larger extent low-earnings immigrants. For instance, in LSIC relative to IMDB average earnings of immigrants predicted to outmigrate in the following wave are estimated as being lower by

[8] Outmigration is defined with respect to income-tax filling. While it is compulsory to file in Canada, we report evidence of non-compliance and thus of mismeasured outmigration.

[9] The approach is based on the work of Hausman, Abrevaya and Scott-Morton (1998) who introduce and estimate misclassification probabilities in binary choice models.

20.6% in 2001 and by 13.6% in 2003, on average for all the cases considered. Therefore, the survey-based estimates considerably overestimate employment and earnings for immigrants predicted to remain in Canada from that cohort, particularly for the last two waves, as well as sizeably overestimate outmigration.

We further find that the model estimated using survey data predicts unsuccessful immigrants (in terms of their unobserved characteristics) are more likely to outmigrate. This is in stark contrast to the estimates derived from the administrative data which predict the exact opposite, *i.e.* successful immigrants (based on their unobservable characteristics) are more likely to outmigrate. Our estimates further suggest that attrition in survey data is dominated not so much by outmigration, but rather by unsuccessful immigrants leaving the panel following a move elsewhere in Canada in search of better outcomes. In contrast, the estimates based upon the administrative data suggest the exact opposite: sample attrition is primarily due to outmigration. Yet, it is important to stress that controlling for measurement error in outmigration partially reduces the gaps in predicted outcomes between the two data sources. Thus, the model we propose may prove useful if only panel survey data is available to investigate the economic performances of immigrants. Nevertheless, sizable gaps remain, suggesting that the inference derived from the administrative data is likely more reliable.

Our main counterfactual analysis seeks to identify which components of the data generating processes are responsible for such diverse outcomes. We do so by sequentially replacing in turn each of the three processes (employment, earnings, outmigration) and two associated covariance matrices estimated from the LSIC with corresponding parameters estimated from the IMDB. Not surprisingly, we find that the differences in predicted labour market trajectories are mainly driven by the outmigration process. In particular, the gaps in predicted employment outcomes are significantly reduced when we replace the outmigration process of the LSIC with that of the IMDB. Proceeding likewise with the earnings or employment processes leads to comparatively smaller reductions in predicted outcomes gaps. The simulation results thus suggest that the differences in predictions using survey and administrative data are largely driven by difficulties in tracking immigrants over time and by not properly accounting for non-random attrition that is unrelated to outmigration in the survey data. It follows that the value of administrative over survey data stems mostly from their capacity to better measure true outmigration.

The remainder of the paper is structured as follows. Section 2 describes the data sources used in the paper. Section 3 presents the econometric model we fit to the data. Section 4 presents the main results of the paper. Section 5 concludes.

2 Data description

This paper uses two data sources. The first is the Longitudinal Survey of Immigrants to Canada (LSIC), a comprehensive longitudinal survey administered by Statistics Canada. The LSIC is composed of three biennial waves during 2001–2005 and captures information about the labour market performance of immigrants, as well as their socio-economic characteristics and background information. The second is the Longitudinal Immigration Database (IMDB). The IMDB contains administrative data from Immigration, Refugees and Citizenship Canada (IRCC), derived from the Field Operations Support System (FOSS), and merged with data from the Canadian Revenue Agency (income tax files); it covers all immigrants who have landed since 1980 and who have filed at least one tax return since 1982.

2.1 Sample Selection

The LSIC focuses on immigrants who arrived in Canada during the period corresponding to October 2000 – September 2001.[10] It is also restricted to immigrants aged 15 or older, and to those who submitted their permanent resident applications from outside Canada through a Canadian mission abroad. The latter exclusion was done in order to focus the survey on newcomers to Canada, since those who landed from within Canada might had already spent a considerable amount of time inside the country under a temporary-resident status.[11] The initial LSIC sample was drawn from the FOSS administrative database. The survey was conducted in three waves with an approximate two-year interval between waves. In particular, the first wave took place between April 2001 – May 2002, the second from December 2002 until December 2003 (with a few interviews in the beginning of 2004), and the third and final wave was conducted from November 2004 until November 2005. The interviews were thus conducted approximately six months, two years and four years upon arrival in Canada.

[10] "Arrival" in the context of this study corresponds to his/her landing as a permanent resident.

[11] Despite this restriction, as we discuss below, some immigrants who applied from outside Canada may still possess some experience of living and working in Canada prior to their landing as permanent residents (though not necessarily immediately before the landing). We take into account this information in the estimations.

The IMDB contains landing information from the FOSS administrative database dating to 1980. It is combined with income-tax declaration data over 1982–2013 (the so-called T1 Family File). It includes information on individual characteristics at landing (education, skill set, origins, age, *etc.*) and all filed income-tax declarations in Canada prior to or after their landing as permanent residents. In order to ensure comparability to the LSIC sample, we restrict our analysis to individuals who arrived between October 2000 – September 2001 and who filed their permanent resident applications through a Canadian mission abroad. Since the LSIC is also restricted to immigrants who were present in Canada at the time of the first-wave interviews, we limit our IMDB sample to those who filed their tax declaration for 2001 and were Canadian residents for tax purposes in that year (*i.e.*, were physically present in Canada, not filing their tax declarations from abroad). Therefore, those who landed over the aforementioned periods but who were not living in Canada in 2001 are excluded from both LSIC and IMDB samples.

We restrict our target population in both samples to males who arrived in Canada under the economic class programs, including both principal applicants and dependants. We next distinguish the economic programs between Business and Non-business classes. Business-class immigrants include those who landed under the categories of Entrepreneurs, Self-employed, Investors or other business-class programs. Non-business class immigrants consist of Skilled Workers, such as those landed under the Federal Skilled Worker (FSW), Canadian Experience Class (CEC), and Provincial/territorial Nominees (PNP) categories. This distinction is in line with the categorization used by IRCC. We further restrict our sample to those aged 25-54 at landing. Additionally, we exclude from the LSIC individuals with imputed wages and start or end dates of employment, as well as those with missing information on earnings. In both datasets, we further exclude individuals who ever reported self-employment income for the years 2001, 2003, and 2005, as well as immigrants who ever resided in the Atlantic provinces (Prince Edward Island, Nova Scotia, New Brunswick, Newfoundland and Labrador), Yukon, Nunavut or the Northwest territories during those years. The latter restriction is imposed due to small samples sizes in LSIC.

Our sample selection strategy aims at matching the LSIC and IMDB samples as closely as possible. However, unavoidable minor differences may remain. Indeed, the IMDB database contains information on individuals who file income-tax declarations in Canada. While filing a tax declaration is mandatory, some might choose not to file if they have no income and are

not eligible to any benefits or to refundable tax credits, or if they are engaged in non-official employment. Given that we restrict our sample to male economic immigrants, we conjecture that potential discrepancies are likely very small.[12] We also conjecture that if a person does not declare his or her income, it is also unlikely that a similar person would willingly report his or her income in the LSIC questionnaire.

2.2 Dependent Variables: Earnings, Employment, Outmigration

2.2.1 Earnings

The IMDB contains information on total annual earnings from employment. We calculate the total wage earnings as the sum of earnings declared in T4 slips and other declared employment income. T4 slips are issued by employers and include all paid-employment income, such as wages, salaries and commissions, before deductions (excluding any self-employment income). Other employment income includes any taxable receipts from employment other than wages, salaries and commissions. These might include tips, gratuities or director's fees that are not reported on a T4 slip, as well as some other components. Tax files do not contain any information regarding hours worked, full- or part-time work, or period of employment within a year. Therefore, we focus our analysis on annual earnings from employment. The annual wage earnings are then adjusted for inflation using the consumer price index (CPI) based in 2011 $ CAD. We also prorate them by the number of days present in Canada for the landing year, using the exact date of landing.

In order to get comparable LSIC labour market information, we first calculate annual wage earnings at each job held during a year by multiplying weekly wage and number of weeks worked. Weekly wages are self-reported and calculated by Statistics Canada from different declared wage types *(e.g.* hourly, weekly, annual, *etc.)* and hours worked.[13] Job duration is derived from the start and end dates of a job. Second, we aggregate wage earnings from

[12]Using IMDB data, we calculated that among male economic-class immigrants who landed in 2000–2001, around 7.7% could not be linked to a tax declaration in any year following their arrival, up to and including 2013. This might include four categories of immigrants: 1) immigrants who landed but did not stay in Canada following landing; 2) immigrants for whom it was impossible to find a match between IRCC and CRA information (the merge is done by Statistics Canada using name, landing date, and other observed characteristics); 3) immigrants who were present in Canada but intentionally have never filed taxes; and 4) immigrants who died in the first year of their arrival.

[13]The wage question in the LSIC is the following: "In this job, what is/was your wage or salary before taxes or other deductions?" Therefore, the wage measure in the LSIC does correspond to the measure of wage earnings reported in tax declarations in the IMDB.

all jobs within each calendar year. As with the IMDB data, we adjust wages using the CPI based in 2011 $ CAD. Third, we prorate total wage earnings by the number of days present in Canada in a given year until the interview date: to do so we use information on the exact date of landing and exact interview dates. For instance, if an immigrant was interviewed in November, we calculate his total earnings income for January-November and then prorate it to get an estimate of the annual income, assuming that his income in December would be similar to his average income over previous months.

2.2.2 Employment

For both the LSIC and the IMDB, an individual is considered employed if his or her annual wage earnings are at least 12,000 $ CAD. Thus, the "unemployed" status in the context of this study includes those who do not work and those who work few hours yearly or whose earnings are relatively modest. The threshold corresponds to the annual earnings of a worker employed full-time at the minimum wage.[14] Additionally, according to the LSIC data, almost all those who reported working more than 30 hours per week also reported earnings of at least 12,000 $ CAD annually.[15]

2.2.3 Attrition and Outmigration

To construct the attrition measure for the LSIC sample, we use information on whether an individual is observed in a subsequent period. Thus our attrition variable is equal to one in 2001 if an individual was interviewed in 2001 but not in 2003, it is equal to zero in 2001 and to one in 2003 if he was interviewed in both 2001 and 2003 but not in 2005, and finally it is equal to zero in both 2001 and 2003 if an immigrant was interviewed in all three waves.[16] In

[14]The minimum wage in Canada in 2001, 2003 and 2005 was 8.65CAD, 8.43CAD and 8.54CAD, respectively (expressed in 2014 CAD; see http://www.statcan.gc.ca/pub/11-630-x/2015006/c-g/desc1-eng.htm). Using 1560 as the level of hours worked in a full-time job in a year (30 hours a week times 52 weeks) and converting to 2011CAD using total CPI, we get the total annual earnings corresponding to the full-time job at a minimum wage at the level of 12,000-13,000 CAD over 2001-2005.

[15]All results presented in the paper were replicated using a 1,000 $ CAD cutoff as a proxy for being employed/unemployed. They were found to be robust to the selection of the cutoff. The results are available upon request. This suggests that the cutoff used to code employment does not play a significant role in our analysis.

[16]Given that during the second and third waves a small fraction of immigrants were interviewed earlier, namely in 2002 and 2004 instead of 2003 and 2005, respectively, and thus that we do not have information on their labour market outcomes for the second and third periods, the attrition in the preceding period is considered still to be zero, since these people remained in the country for the subsequent interview despite the absence of the labour market information.

our LSIC sample, attrition status is equal to one for 32.9% and 18.9% of immigrants in the first and second periods, respectively. As we show next, these attrition rates are much larger than those observed in administrative data. Naturally, the attrition status is unobserved for the third period, because we do not have information on whether an immigrant remained or not in the country after 2005, *i.e.* after the third and last wave of the interviews. The attrition is quantitatively smaller in the second period, but remains important.

For the IMDB, we construct a similar measure of attrition. The attrition status in period t is thus equal to one if a person does not file a tax declaration in the subsequent period $t+1$. Moreover, if an individual does not file a tax declaration in 2003 but does so in 2005, the attrition status is nevertheless set to one in $t = 2001$. We do this to mimic as closely as possible the sampling framework of the LSIC. The resulting attrition rates in $t = 2001$ and $t = 2003$ are 5.0% and 6.8%, respectively, thus six and tree times smaller than those observed in LSIC.

Attrition in LSIC includes both individuals who outmigrated and who remained in Canada but could not be reached for the second or third interviews. This includes any temporary departures from Canada occurred in the survey years. Attrition in the IMDB occurs whenever someone leaves the country or stops filing an income-tax declaration while still residing in Canada. To mimick LSIC the constructed attrition measure in IMDB also includes those who temporarily departed or temporarily stopped filing taxes.[17] As mentionned in the introduction, the analysis of Bryan, Chowdhury and Mobarak (2014) suggests that the vast majority of IMDB immigrants not filling in a given year is explained by outmigration rather temporarily not filling taxes while present in the country in a given year. What is more, the lower levels of attrition in our IMDB sample suggests that it better tracks those who remain in Canada. Given that the proportion of immigrants with positive attrition is not negligible in either samples, it is important to account for potential non-random attrition when investigating labour market performances, even more so when using survey data.

2.3 Explanatory Variables

The LSIC and IMDB samples contain the following set of similarly defined individual characteristics: age, province of residence, immigration class, skill level as assessed by IRCC, edu-

[17] Section 4.4 discusses estimations for IMDB using an alternative attrition measure that focuses on permanent outmigration and thus takes into account all the available information in administrative records on the immigrants' presence in the country during other years. All our results are largely unaffected by this change.

cational level, country of birth and country of last permanent residency, presence in Canada prior to landing, number of other immigrants in the family, and non-employment income.[18] Some characteristics are derived from the FOSS records in both the LSIC and IMDB; others are self-reported in the LSIC, while the same characteristics come from the administrative records in the IMDB. Additionally, we merge several indicators related to the country of birth, which we derive from external sources. Namely, we add information on *per capita* GDP, population density, unemployment and emigration rates observed in the country of birth of an immigrant, its distance to Canada, and difference in mean temperatures between the country of birth and Canada. Appendix A provides further details on these additional variables.

2.4 Comparison of the LSIC and IMDB samples

Table 1 lists summary statistics from the LSIC and IMDB samples. Three main observations are worth noting from this table. First, as noted previously, the attrition rates are much higher in the LSIC than in the IMDB. Yet, attrition rates in the IMDB are not negligible. This suggests that attrition may be non-random. Therefore, correcting for potential biases due to non-random attrition is important for both datasets, although primarily so when using the LSIC sample.

Second, individual characteristics are similar across the two samples at baseline (2001). This is particularly true for variables drawn from the FOSS: immigration class, country of birth and skill level. There are some differences in variables that are self-reported in the LSIC and derived from tax files in the IMDB, but these differences are relatively small: province of residence, earnings, non-employment income.

Third, sizable differences between the two samples appear in the two periods following the baseline year. For instance, the fraction of principal applicants in the non-business class increases in the LSIC and declines in the IMDB by 2005; the fraction of immigrants with higher education declines by 2.1 percentage points in the LSIC, while it only declines by 0.2 percentage points in the IMDB; the proportion of Ontario residents declines to below 50% in the LSIC – from 57.8% to 49.8%, while this proportion in the IMDB decreases only from 57.9% to 56.6%; the proportion of Quebec residents increases in the LSIC from 20.9% to 25.2%, while it slightly declines in the IMDB from 17.7% to 17.0%; the fraction of immigrants originating from Asia, Australasia and the Pacific declines in the LSIC by 5 percentage points, while it increases by

[18] Table B1 in Appendix B describes the source of information for these variables in both datasets.

less than 1 percentage point in the IMDB.

Importantly, the LSIC data also reveals faster employment and earnings growth than what is reported in the IMDB. In particular, earnings increased by as much as 34% between 2001-2005 in the LSIC, but only by 22% in the IMDB. It may be argued that non-random attrition is partly driving these compositional changes and economic outcomes.

3 The Model

Our model is defined over three periods (data waves), denoted t = 1, 2, 3, corresponding to years 2001, 2003, and 2005. Let $i = 1, 2, ..., N$ denote individuals in a given sample. Each period we observe (full-time equivalent) employment status (e_{it}), log of annual earnings (w_{it}), and whether the immigrant leaves the panel in the following period, a binary variable denoted r_{it}. Employment is assumed to be generated by the following latent process:

$$e_{it} = \begin{cases} 1, \text{ if } e_{it}^* \geqslant 0 \\ 0, \text{ if } e_{it}^* < 0 \end{cases}, \quad e_{it}^* = \mathbf{x}_{it}^{e'}\boldsymbol{\delta} + \eta_i^e + \varepsilon_{it}^e \,, \tag{1}$$

where $\mathbf{x}_{it}^e$ represents a vector of covariates, $\boldsymbol{\delta}$ denotes a vector of unknown parameters, η_i^e represents individual-specific time-invariant unobserved characteristics such as unobserved individual ability, and ε_{it}^e captures time varying stochastic shocks. We allow for dynamic dependency by including lagged employment status e_{it-1} as a covariate.

We model full-time equivalent earnings using the following specification:

$$w_{it} = \mathbf{x}_{it}^{w'}\boldsymbol{\beta} + \eta_i^w + \varepsilon_{it}^w, \tag{2}$$

where $\mathbf{x}_{it}^w$ represents observable individual characteristics, $\boldsymbol{\beta}$ denotes a vector of unknown parameters, η_i^w represents individual-specific time-invariant unobserved characteristics, and ε_{it}^w captures stochastic shocks to annual employment earnings. Full-time equivalent earnings are only observable when $e_{it} = 1$.

Outmigration in period t is denoted by the (imperfectly observed) binary variable r_{it}^0, where $r_{it}^o = 1$ whenever an individual leaves the host country in the following period. In practice, we do not observe r_{it}^0 but rather observe sample attrition, r_{it}. An individual who outmigrates

($r_{it}^0 = 1$) also leaves the sample ($r_{it} = 1$). An individual who remains in the panel necessarily does not outmigrate, hence $r_{it}^0 = 0$ when $r_{it} = 0$. On the other hand, some may leave the sample in the following period ($r_{it} = 1$) yet remain in the country ($r_{it}^0 = 0$). We discuss below how we handle partial observability of r_{it}^0.

True outmigration is assumed to be determined by the following latent process :

$$r_{it}^o = \begin{cases} 1, \text{ if } r_{it}^{o*} \geqslant 0 \\ 0, \text{ if } r_{it}^{o*} < 0 \end{cases}, \quad r_{it}^{o*} = \mathbf{x}_{it}^{o'}\boldsymbol{\gamma} + \eta_i^r + \varepsilon_{it}^r, \tag{3}$$

where $\mathbf{x}_{it}^o$ is a vector of covariates including lagged employment status e_{it-1}, $\boldsymbol{\gamma}$ is a vector of unknown parameters, while η_i^r denotes individual-specific time-invariant unobserved characteristics. Finally, ε_{it}^r captures time varying stochastic shocks.

The unobserved heterogeneity parameters $\{\eta_i^1, \eta_i^2, \eta_i^3\}$ are assumed to be jointly normally distributed with mean and covariance matrices as follows:

$$\begin{Bmatrix} \eta_i^e \\ \eta_i^r \\ \eta_i^w \end{Bmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_\eta = \begin{pmatrix} \sigma_{\eta_e}^2 & \rho_{\eta_r\eta_e}\sigma_{\eta_e}\sigma_{\eta_r} & \rho_{\eta_e\eta_w}\sigma_{\eta_e}\sigma_{\eta_w} \\ \rho_{\eta_r\eta_e}\sigma_{\eta_e}\sigma_{\eta_r} & \sigma_{\eta_r}^2 & \rho_{\eta_r\eta_w}\sigma_{\eta_r}\sigma_{\eta_w} \\ \rho_{\eta_e\eta_w}\sigma_{\eta_e}\sigma_{\eta_w} & \rho_{\eta_r\eta_w}\sigma_{\eta_r}\sigma_{\eta_w} & \sigma_{\eta_w}^2 \end{pmatrix} \right\} \tag{4}$$

The stochastic terms $\{\varepsilon_i^e, \varepsilon_i^r, \varepsilon_i^w\}$ are assumed to be independent and identically normally distributed over time with mean and covariance matrices as below:

$$\begin{Bmatrix} \varepsilon_{it}^e \\ \varepsilon_{it}^r \\ \varepsilon_{it}^w \end{Bmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_\varepsilon = \begin{pmatrix} \sigma_{\varepsilon_e}^2 & \rho_{\varepsilon_r\varepsilon_e}\sigma_{\varepsilon_e}\sigma_{\varepsilon_r} & \rho_{\varepsilon_e\varepsilon_w}\sigma_{\varepsilon_e}\sigma_{\varepsilon_w} \\ \rho_{\varepsilon_r\varepsilon_e}\sigma_{\varepsilon_e}\sigma_{\varepsilon_r} & \sigma_{\varepsilon_r}^2 & \rho_{\varepsilon_r\varepsilon_w}\sigma_{\varepsilon_r}\sigma_{\varepsilon_w} \\ \rho_{\varepsilon_e\varepsilon_w}\sigma_{\varepsilon_e}\sigma_{\varepsilon_w} & \rho_{\varepsilon_r\varepsilon_w}\sigma_{\varepsilon_r}\sigma_{\varepsilon_w} & \sigma_{\varepsilon_w}^2 \end{pmatrix} \right\} \tag{5}$$

All random components of the model are assumed to be independent of the covariates. For identification purposes, the variances of the unobserved stochastic terms entering the equations for employment and outmigration are set to 1: $\sigma_{\varepsilon_e}^2 = 1$ and $\sigma_{\varepsilon_r}^2 = 1$. The other parameters of both covariance matrices are treated as free parameters.

All three equations of our model (employment, outmigration, earnings) include lagged em-

ployment status, e_{it-1}, as an explanatory variable, thus raising the issue of initial conditions. This problem arises when the initial value of the lagged endogenous explanatory variable is stochastic and possibly non-random. In our case, the cohort of immigrants we investigate had no immediately prior labour market experience in Canada at landing. Consequently, we set $e_{it-1} = 0$ in 2001 and do not need to adjust the model in any particular way.[19]

The above model is incomplete insofar as attrition status, r_{it}, is observed not outmigration, r^0_{it}, *per se*. We address this measurement issue by extending an approach due to Bellemare (2007). As above, let r_{it} denote attrition, a potentially mismeasured indicator of outmigration. Given our panel is defined over three periods, r_{it} is observed for the first and second periods and is treated as unobservable in the third. Despite the fact that r_{i3} is observable in the IMDB sample, we treat it as unobservable to maintain symmetry with the LSIC sample. From the law of total probability we get

$$\begin{aligned} P(r_{it} = 1|\mathbf{x}^o_{it}) &= \alpha_t + (1 - \alpha_t) \cdot P(r^o_{it} = 1|\mathbf{x}^o_{it}) \, ; \\ P(r_{it} = 0|\mathbf{x}^o_{it}) &= (1 - \alpha_t) \cdot P(r^o_{it} = 0|\mathbf{x}^o_{it}) \, , \end{aligned} \tag{6}$$

where $P(r^o_{it} = 1|\mathbf{x}^o_{it})$ denotes the true outmigration probability given observable covariates $\mathbf{x}^o_{it}$, and α_t is a measurement error. Specifically, α_t captures the probability that true outmigration differs from observed attrition. The probability $P(r^o_{it} = 1|\mathbf{x}^o_{it})$ is derived from the latent process for r^o_{it} presented above. We treat (α_1, α_2) as fixed parameters to be estimated along with the other model parameters. Both (α_1, α_2) are thus assumed independent of the covariates $\mathbf{x}^o_{it}$. The latter assumption is traditionally maintained in order to identify models with measurement error in which the endogenous variable is discrete (see Lewbel, 2000). *A priori*, we expect the estimates of (α_1, α_2) to be smaller when the model is estimated with the IMDB sample since individuals are better tracked over time, for instance when they move within Canada, and attrition is thus likely a less noisier indicator of outmigration.

We also considered two variations of the model above. First, we estimated a restricted specification which does not incorporate misclassification probabilities (*i.e.*, setting $\alpha_1 = 0, \alpha_2 = 0$). The substantive differences between LSIC and IMDB estimates and predictions we highlight in sections 4.2 and 4.3 below are exacerbated when doing so. Second, our main IMDB sample

[19] Our cohort of immigrants landed in Canada in October 2000 at the earliest. See Section 2 for details.

was constructed to mimic as closely as possible the sampling framework of the LSIC survey sample which treats attrition as permanent. This was enforced by setting $r_{it} = 1$ whenever an individual failed to file a tax return in the following period, regardless of whether a tax return was filed in subsequent years. We also estimated the model on the IMDB sample using a definition of attrition that reflects permanent attrition/outmigration by setting $r_{it} = 1$ when no tax returns were filed in all following years. This measure of attrition thus abstracts from any temporary absences of filed tax declarations while being present in Canada as well as from any temporary departures from Canada. We find that the estimated values of α_1 and α_2 are smaller in this case. Section 4.4 discusses these variations of the model in more details.

The model is estimated by simulated maximum likelihood. Appendix C provides further details about the maximum likelihood function and estimation procedure. In addition to the normalization of $\sigma^2_{\varepsilon_e}$ and $\sigma^2_{\varepsilon_r}$, the identification of the empirical model requires some exclusion restrictions. To that end, the following are deemed appropriate identifying variables in the attrition equation: a dummy variable indicating whether the country of last permanent residence is the same as the country of birth; the distance from the country of birth to Canada (in log); the difference in mean temperatures between the country of birth and Canada (in °C); the emigration rate of tertiary education graduates observed in the country of birth; the contemporaneous unemployment rate in the country of birth; and the number of other immigrants in the family.[20] Likewise, the log of non-employment income is included in the employment equation only.[21]

4 Estimation Results

4.1 Parameter estimates

Tables 2a, 2b, and 2c present the parameter estimates from the described above model for the outmigration, employment and wage equations, respectively. Table 2d lists estimated parameters for the covariance matrices of unobserved terms. The results are presented separately for the survey (LSIC) and the administrative (IMDB) data.[22]

[20] Appendix A describes in detail variables capturing country of birth characteristics.

[21] The exclusion restrictions are more stringent than necessary. In theory, identification can be established if all but one continuous covariate in $\mathbf{x}^0_{it}$ affects α_t. We did not pursue this approach since we have few continuous variables in the data.

[22] As mentioned earlier, the IMDB sample mimicks the LSIC sample and the attrition measure in IMDB mimicks the LSIC sampling framework.

Outmigration equation

The parameter estimates in Table 2a suggest that the determinants of outmigration differ significantly across the two samples. In many cases, statistically significant parameters differ in sign, in magnitude or both. Importantly, principal applicants under the business class in the LSIC sample are found to have a higher probability of outmigrating relative to the dependent applicants under economic class while no such evidence is found in the IMDB data. In contrast, principal applicants under the non-business economic class are found to have a significantly lower probability of outmigrating using IMDB data, while this coefficient is four times smaller and not statistically significant in LSIC. We also find opposite results concerning skill levels. In particular, managers and professionals (levels "0" and "A") are found to be more likely to leave based on IMDB data, but not in LSIC data. In contrast, immigrants with lower skill levels (levels "B", "C", "D") are found to less likely outmigrate using LSIC data, but not so using IMDB data.[23] In both samples, immigrants to Ontario are found to be more likely to leave than those who landed in British Columbia while those from Quebec are found to be less likely to do so only in the LSIC sample. This is perhaps related to the fact that immigrants to Quebec are less mobile due to the language barrier.

The characteristics of the source countries yield relatively similar results across samples, with a few exceptions. According to both datasets, immigrants coming from countries with higher GDP *per capita* and lower unemployment rates are more likely to outmigrate, as expected. Immigrants who had previous international moves, namely immigrants for whom the country of last permanent residence was different from the country of birth, are also more likely to outmigrate. Both sets of parameters indicate that the probability of leaving is no higher for immigrants originating from World area 1 than those from the USA and Europe. In contrast, while LSIC data suggest that immigrants coming from World area 2 are more likely to outmigrate than immigrants from the USA and Europe, the opposite is found in the IMDB data. Interestingly, mean temperature differences between the source country and Canada (usually positive) are found to increase the probability of leaving in both data sets. Higher emigration rates of tertiary education population in the source country are found to negatively affect outmigration probability in both cases, but the coefficient is statistically significant only in IMDB.

The last panel of the table focuses on past employment and measurement errors. Not sur-

[23] See Table 1 and B1 for details on the used explanatory variables.

prisingly, those who were employed in the previous wave are found to be much less likely to have left according to both the LSIC and the IMDB. Further, a comparison of the α_t parameters shows that the attrition indicator used in the LSIC data is considerably more contaminated by measurement errors than that used in the IMDB data despite the fact that the latter was drawn so as to mimic the former. This result was expected given significantly higher attrition rates in LSIC data. According to the LSIC parameter estimates, the estimated probabilities of exiting the survey and staying in the country are 11.3% and 11.6% for the 2001 and 2003 waves, respectively. These estimates mirror those of Bellemare (2007) who estimated a corresponding misclassification probability of 10.9% in the immigrant sample of the German Socio-Economic Panel. The estimated missclassification probabilities are also statistically significant in the administrative sample but they are almost ten times lower than that of LSIC (corresponding estimated probabilities are 1.6% and 1.4%, respectively for 2001 and 2003). Given that attrition rates in IMDB are 5.0 and 6.8% in 2001 and 2003, the estimation results imply that the vast majority of sample attrition in the administrative data is genuine outmigration.[24]

Based on the estimates of the LSIC model, the predicted outmigration probabilities for the entire cohort of immigrants in LSIC are estimated at 24.2% and 7.5% in 2001 and 2003, respectively. The predicted outmigration probabilities for the IMDB sample for 2001 and 2003 are 3.6% and 5.2%, respectively. Higher estimates for missclassification probabilities α_t in LSIC reflect higher random attrition of immigrants between waves in survey versus administrative data. Higher predicted outmigration probabilities in LSIC versus IMDB are indicative of an additional non-random attrition from the LSIC sample between waves, which is particularly pronounced for the period between the first and second interviews. Below, we discuss the differences in predicted attrition in more details.

Employment equation

Table 2b reports the parameter estimates of the employment equation using both samples. Nearly all of them are statistically significant, most bear the same sign and many are more or less of the same magnitude. There are a few noteworthy exceptions, though. For instance, according to both data sources, non-business class immigrants are more likely to be employed and the business class much less so relative to the reference category of dependents under the

[24]This is in line with Aydemir and Robinson (2008) who, by comparing immigrants' retention rates using Census and IMDB data, suggest that non-filing behaviour (for four consecutive years) in administrative tax data is mostly associated with absence of immigrants from Canada rather than with being in the country and not filing.

economic class; however, the magnitude of these effects is twice as large in the LSIC. Additionally, we find conflicting results concerning skill levels and education. In particular, higher education is associated with a lower probability of employment in the LSIC sample while the opposite holds in the IMDB sample. The conflicting results in the employment equation mirror those in the attrition equation. The estimated parameters for year dummies and for dynamic dependence are quite similar across two datasets. Namely, the lagged employment status variable is found to significantly increase the employment probability in both cases, with a slightly larger coefficient estimated in LSIC. Coefficients for yearly dummy variables have a similar magnitude in both data sets and demonstrate an increasing employment probability of immigrants with their tenure in Canada.

Earnings equation

Unlike the two previous equations, the parameter estimates of the earnings equation in Table 2c are remarkably similar across the two samples. On the whole, the parameter estimates are consistent with traditional human capital equations: earnings increase with age (at a decreasing rate), skill level and education. Consistent with the literature for Canada, immigrants to Ontario and the Prairies earn slightly more than those who landed in British Columbia. Those who landed in Quebec, on the other hand, earn slightly less. Immigrants' tenure in Canada, represented by yearly dummy variables, increases earnings at a similar magnitude in both cases. The coefficient for the lagged employment status is found to be positive and significant in both datasets, but its value in IMDB is twice as high as in LSIC data.

Correlation structures

Table 2d presents the estimated covariance matrices (see equations (4) and (5)). The top panel focuses on the contemporaneous error terms (equation (5)). The parameter estimates differ considerably between samples. Those of the LSIC sample are at least halved relative to those of the IMDB data. In particular, the conditional variance of the earnings equation is much larger in the latter, despite the fact that the slope coefficients derived from the two samples for this equation are very similar. Both samples yield a negative correlation between employment and outmigration, as expected. In addition, both samples also yield a negative, yet smaller in absolute value, correlation between earnings and outmigration. And for both these correlations, their absolute values in the LSIC sample are more than two times lower relative to IMDB.

Therefore, negative shocks to employment or earnings increase outmigration probabilities to a larger extent according to the administrative than the survey data. Surprisingly, the correlation between employment and earnings is found to be negative in the LSIC sample but to be statistically insignificant in the IMDB sample.

The bottom panel of the table focuses on individual effects, $\eta_i^j, j = e, r, w$ (equation (4)). The estimated variance of the individual effects in the outmigration equation is smaller in the LSIC data. On the other hand, the variances in the employment and earnings equations are nearly identical. Yet, the striking feature of the panel concerns the correlation coefficients $\rho_{\eta_{e,r}}$ and $\rho_{\eta_{r,w}}$. While they are highly statistically significant in both samples, they bear opposite signs. Namely, in the LSIC data it is found that the unobserved individual characteristics are negatively correlated between the outmigration and employment equations, as well as between the outmigration and earnings equations. This is consistent with the fact that the attrition in the LSIC is severely ill-measured. Despite the fact that the model attempts to account for this mismeasurement, those who leave the sample may have done so due to poor earnings or employment outcomes. In other words, these results suggest that the attrition that is not related to outmigration is not random in LSIC, but instead is linked to lower earnings and employment outcomes. For instance, moving within Canada is likely linked to one's unsatisfactory employment and earnings. This, in turn, leads to sample attrition not outmigration *per se*. Mismeasurement of outmigration is much less an issue in the IMDB data, and our estimates suggest that attrition reflects true outmigration to a large extent. Thus the positive correlations between outmigration and employment on the one hand, and outmigration and earnings on the other hand, more likely reflect selection into outmigration from high-earners and highly employable immigrants. This is entirely consistent with the findings of Damas de Matos and Parent (2019) according to which young and highly-skilled immigrants to Canada are the most likely to outmigrate to the United States. These results also suggest that a large share of the attrition in the LSIC data is non-random and unrelated to outmigration. The fact that true outmigration is difficult to identify when using survey data has important policy implications. Indeed, survey data may significantly bias measured labour market performances due to its inability to properly account for sample attrition. We investigate this issue in what follows by contrasting the predicted labour market trajectories using the two sets of estimates.

4.2 Predicted Outcomes

4.2.1 Employment Transitions

Table 3 presents the predicted distribution of possible employment and outmigration outcomes based on the estimates from the IMDB sample. The predictions are computed for 16 different subsets of individuals according to province of residence (Ontario, Québec, British Columbia, Prairies), age in 2001 (25 or 35) and education level (post-secondary or higher education). The columns capture the 14 potential transitions between employment, unemployment and outmigration over the three years spanned by our analysis. For instance, the first column focuses on the permanently employed (labelled as "(**e**,**e**,**e**)"). The second column corresponds to the employed-employed-unemployed transition (labelled as "(**e**,**e**,**n**)"). Outmigration is indicated using the dot symbol ("·"). For example, the last column corresponds to the unemployment status followed by outmigration (labelled as "(**n**,·,·)").[25]

We find that the modal trajectory for all conditional subsets is permanent employment, with the predicted shares ranging from 27.2% to 57.2% (40% on average). The maximum value occurs in the Prairies which exhibit the highest predicted employment rates across all provinces, irrespective of age and education.[26] The predicted shares for the modal trajectory are largely insensitive to education. The second most important transition concerns those unemployed during the first year but employed in the following two years. The estimated shares average 22%, with little variation across the different subsets. Finally, predicted outmigration rates are relatively low. In particular, the row sums of the last two columns provide an estimate of the shares of individuals who outmigrate after 2001, whether originally employed or not. These fall below 5% for all subsets.

Our main interest lies in the contrast between the predicted outcomes based on the LSIC parameter estimates with those of the IMDB sample. To this end, the same set of simulations were conducted using the parameter estimates of the LSIC sample and the mean responses were contrasted with those of Table 3.[27] These *level differences* (in percentage points) are reported in

[25] Predictions for each conditional subset were generated by simulating 1 million trajectories for 2001, 2003, and 2005 using draws from the corresponding estimated distributions of unobservables. Specifically, for each trajectory we draw the triplet $(\eta_i^r, \eta_i^e, \eta_i^w)$ and random shocks $(\varepsilon_{it}^r, \varepsilon_{it}^e, \varepsilon_{it}^w)$ for $t = 1, 2, 3$. Employment, outmigration, and earnings (when relevant) are then predicted using the model structure presented in Section 3. The predicted trajectories are averaged out so as to derive the distributions of potential trajectories.

[26] This is in line with the national labor force statistics: the Prairies had the highest employment rates among the provinces during 2001-2005. This period coincides with the beginning of the oil-price boom.

[27] The predicted shares using the LSIC database are available upon request.

Table 4. For ease of reading, negative and positive differences of 2 or more percentage points are highlighted in red and yellow, respectively. Negative (positive) differences imply that the mean shares predicted from the LSIC parameter estimates are lower (higher) than those based on the IMDB estimates.

Consistent with the higher attrition rates observed in the LSIC, we find that the probabilities of outmigration from Canada after the first sample year (last two columns) are predicted to be significantly higher when using survey data. The estimated differences are sizable for all subsets, and in particular for the highly educated who landed in British Columbia or Ontario. In the latter cases, predicted outmigration after a single wave is more than 25 percentage points higher in survey data. Given that predicted outmigration rates in the administrative database are very low, these differences represent an increase of up to 600% relative to those of the IMDB sample. On average, the proportion of outmigrants after both waves is predicted to be 24.7% in the LSIC and only 7.1% in the IMDB.[28] Because predicted shares sum to 100%, the excess predicted outmigration in the survey database implies a redistribution of predicted shares. Hence, it is found that survey estimates significantly underestimate the proportion of immigrants from that cohort being permanently employed (first column), and more so for the highly educated. For example, the estimated shares in Ontario and British Columbia are more than 7 percentage points lower when using survey estimates. The only exception concerns the Prairie provinces for which the LSIC estimates exceed those of the IMDB by slightly less than 3 percentage points for those with post-secondary education. Similarly, survey estimates considerably underpredict the shares of those who are unemployed in the first wave but employed in the following two waves. This holds for all the conditional samples, with 14 out of 16 being underpredicted by more than 2 percentage points, reaching more than 9 percentage points for those who have landed in Ontario. Given that the "(**n**,**e**,**e**)" transition is estimated at approximately 22% in Ontario using the IMDB data (see Table 3), it follows the LSIC predictions are underestimated by as much as 40%. By and large, these results indicate that a larger attrition in survey data leads to a considerable overestimation of the number of outmigrating individuals and to an important underestimation of immigrants' successful economic integration.

The LSIC-based predictions underestimate to a larger extent the proportions of immigrants exhibiting a weak labour market attachment or unable to find employment. In particular, the

[28] These numbers represent the total of six trajectories: "(**e**,**e**,.)", "(**e**,**n**,.)", "(**n**,**e**,.)", "(**n**,**n**,.)", "(**e**,.)", "(**n**,.)".

proportion of immigrants employed in all three waves in the LSIC, "(**e**,**e**,**e**)", is predicted to be 36.7%. Relative to the mean of the IMDB sample at 40%, it is thus underestimated by 8% (and up to 23% for some cases of higher educated immigrants). Likewise, the predicted proportions of trajectories with weak labour market attachment, *i.e.* being employed in only one or only two waves, also are underpredicted in LSIC by 25% and 28%, respectively.[29] Finally, the proportion of immigrants never employed over the three waves is predicted to be 7.2% and 10.3% in the LSIC and IMDB, respectively. This represent an underestimation of over 30% by LSIC.

These simulations, thus, suggest that LSIC-based estimates largely overestimate the number of outmigrants during the first five years following landing. Consequently, they considerably underestimate the proportion of immigrants staying in the country and being continuously employed among the landed cohort, as well as underestimate to an even larger extent the proportion of stayers among this cohort with weaker labour market transitions.

4.2.2 Predicted Employment

Table 5 provides further evidence that employment predictions may be considerably biased when using survey data exhibiting large attrition rates. The left-hand side panel reports predicted employment probabilities of individuals who are predicted to be present in Canada in each respective year. For instance, predicted employment probabilities in 2003 are reported only for those who are predicted to remain in the country after 2001, while predicted employment probabilities in 2005 are calculated only for those who are predicted to stay in Canada until the third wave. Columns 2–4 reports predicted employment probabilities using IMDB estimates, while columns 5–7 show predicted differences between the LSIC and IMDB estimates. According to the IMDB model, the employment rates increase steadily between 2001 and 2005, from more or less 50% to approximately 80%. The LSIC model, on the other hand, underestimates the employment rates of residents in Québec and BC in 2001, and overestimates them in 2003 and 2005 for nearly every subsets. It is important to note that the differences in employment probabilities predicted using the IMDB and the LSIC become more pronounced over time for the majority of cases we consider.

The right hand side of the Table presents the quantitative predicted differences in employ-

[29] Being employed in any two waves, while being present in Canada for all three waves, is represented by the sum of the following three trajectories: "(**n**,**e**,**e**)", "(**e**,**n**,**e**)", "(**e**,**e**,**n**)". Employed in a single wave corresponds to the sum of the following trajectories: "(**e**,**n**,**n**)", "(**n**,**e**,**n**)", "(**n**,**n**,**e**)".

ment outcomes in each reported year for immigrants who are predicted to leave or stay in Canada after that reported year. Since we do not observe attrition and thus do not model outmigration after 2005, these employment probabilities are reported for 2001 and 2003 only. Again, the table reports predicted employment status using IMDB estimates and the difference between simulations based on LSIC and IMDB estimates. According to the IMDB model, those who are predicted to leave Canada in the next period have lower employment rates, ranging on average from 21% in 2001 to 34% in 2003. Once again, the LSIC model overestimates the employments rates by 6.3 percentage points in 2001 and by 3 percentage points in 2003, save for those residing in Québec. The upward bias of the LSIC model also applies to individuals who are predicted to remain in Canada in the next period. Indeed, while the IMDB model predicts average employment rates of 50% and 73% in 2001 and 2003, respectively, the LSIC model overpredicts these rates by 4 and 2.4 percentage points for the same years on average, except for highly educated Québec residents whose employment rates are once again underestimated.

Overall, the predicted employment probabilities using IMDB estimates suggest that outmigrants have lower employment rates in the year preceding outmigration. LSIC also captures this pattern. However, since the LSIC model overpredicts outmigration, and given that outmigrants have lower employment rates than those predicted to remain in Canada, LSIC estimates systematically overpredict the average employment rate of immigrants who remain in Canada with an increasing bias over time.

4.2.3 Predicted Earnings

Table 6 is structured like the previous table: the left-hand side panel focuses on those who are predicted to be in Canada and employed in the reported year, whereas the right-hand side panel distinguishes between those who are predicted to outmigrate or not in the subsequent period. Overall, the IMDB model predicts a sizable increase in earnings for those who remain in Canada between 2001 and 2005, commensurate to the increase in employment reported in Table 5. The LSIC model, on the other hand, overestimates the earnings in 2003 and 2005, but to a lesser extent in the latter case. This is not surprising given that it also overestimates employment for the pool of immigrants predicted to remain in Canada.

According to the IMDB estimates on right-hand side panel, individuals who are predicted to outmigrate after the 2001 wave have higher predicted earnings than those who are pre-

dicted to stay – 37,266$ vs 32,460 $ CAD for movers and stayers, respectively. The converse holds for the 2003 wave, with stayers predicted to have slightly higher average earnings. Differences in expected earnings between LSIC and IMDB are sizable. Indeed, for the 2001 and 2003 waves the differences are as large as -20.6% and -13.6% for outmigrants, whereas corresponding differences for stayers are +4.4% and +9.3%. Clearly, survey data finds stronger negative self-selection into outmigration than administrative data. This is consistent with the estimated covariance matrices of the unobservable components of the the model (Table 2d). Consequently, the LSIC overpredicts the fraction of low performing who leave the country and depicts overoptimistic labour market performance for those who remain in the country.

4.3 Decomposing the Differences Between the LSIC and the IMBD Transitions

Differences in predicted outcomes result from differences in the sample-specific parameter estimates of our three-equation model. We thus conduct additional simulations to investigate the sources of the discrepancies in the predicted outcomes. We use the predictions reported in Table 3 derived from IMDB sample as our benchmark. We contrast these with 30 counterfactual predictions of employment/outmigration trajectories for each of the 16 subsets of immigrants reported in the table. Each counterfactual prediction is based upon the LSIC parameter estimates, replacing in turn selected subsets of parameters with corresponding estimates obtained using the IMDB sample. Through sequential substitutions, we are able to identify the set of parameters estimates who contribute most to the differences in predicted trajectories. To be more specific, let $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, and $\boldsymbol{\beta}$ denote the vectors of parameters for the outmigration, employment, and wage equations, respectively. Also, let $\boldsymbol{\Sigma}_\epsilon$ and $\boldsymbol{\Sigma}_\eta$ denote the covariance matrices of $(\varepsilon_i^e, \varepsilon_i^r, \varepsilon_i^w)$ and $(\eta_i^e, \eta_i^r, \eta_i^w)$. Each sample provides a distinct set of estimates of $(\boldsymbol{\gamma}^d, \boldsymbol{\delta}^d, \boldsymbol{\beta}^d, \boldsymbol{\Sigma}_\epsilon^d, \boldsymbol{\Sigma}_\eta^d)$, where d indexes the data source ($d \in \{I, L\}$) where I and L denote IMDB and LSIC based estimates, respectively (see Tables 2a to 2d for the estimates).

Given the above notation, we compute counterfactual trajectories for various combinations of subsets of parameter estimates from both data sources. For example, we begin by predicting trajectories using $(\boldsymbol{\gamma}^I, \boldsymbol{\delta}^L, \boldsymbol{\beta}^L, \boldsymbol{\Sigma}_\epsilon^L, \boldsymbol{\Sigma}_\eta^L)$, *i.e.* using all LSIC parameter estimates save for the outmigration equation coefficients. This yields a table of predicted trajectories akin to Table 3. We then subtract this counterfactual table from Table 3 which yields a table of similar to Table 4. The exercise above is repeated using $(\boldsymbol{\gamma}^L, \boldsymbol{\delta}^I, \boldsymbol{\beta}^L, \boldsymbol{\Sigma}_\epsilon^L, \boldsymbol{\Sigma}_\eta^L)$. Thus, LSIC-based employment equa-

tion parameter estimates are replaced by those of the the IMDB estimates, all other parameters remaining fixed at their LSIC values. We repeat this exercise for all 30 possible combinations of parameters.

We measure the discrepancies between the benchmark and the counterfactuals using both the sum of absolute and the sum of quadratic deviations of predicted counterfactual trajectories for each of the 16 subsets of immigrants. Table 7 reports the average, minimum and maximum sum of absolute/squared deviations across all 16 subsets of immigrants for the different counterfactual predictions.

The first line of Table 7 reports the mean sum of absolute/squared deviations between the raw LSIC and IMDB trajectories. The respective differences are 38.0 and 302.0, which is considerable. In others words, the two samples yield entirely different labour market stories as previously highlighted in Table 4. We begin by discussing the impact of substituting a single subset of parameters at a time. These simulations are reported in the second panel of the table. Replacing the outmigration LSIC parameters by those of the IMDB sample has a sizable impact on predicted trajectories. We find that the mean absolute/squared deviations fall to 11.9/28.8, a 68.7% and 90.4% reduction relative to the benchmark deviations. Conversely, replacing either employment ($\boldsymbol{\delta}^I$ instead of $\boldsymbol{\delta}^L$), wage ($\boldsymbol{\beta}^I$ instead of $\boldsymbol{\beta}^L$) or covariance ($\boldsymbol{\Sigma}_\epsilon^I$ instead of $\boldsymbol{\Sigma}_\epsilon^L$) parameter estimates has a limited impact on the deviations. On the other hand, substituting the covariance matrices of the individual fixed-effects ($\boldsymbol{\Sigma}_\eta^I$ instead of $\boldsymbol{\Sigma}_\eta^L$) further increases the mean absolute/squared deviations by 10.2% and 15.9% respectively. The remaining panels of the Table 7 investigate the impact of simultaneously changing two or more subvectors of parameter estimates at once (for example predicting counterfactual trajectories using $(\boldsymbol{\gamma}^I, \boldsymbol{\delta}^I, \boldsymbol{\beta}^L, \boldsymbol{\Sigma}_\epsilon^L, \boldsymbol{\Sigma}_\eta^L)$. Significant reductions in mean absolute/squared deviations only occur when parameter estimates of the outmigration equations are substituted in. These results highlight the fact that the outmigration or attrition process is the core element explaining the different predicted employment/outmigration trajectories from survey and administrative data.

Additionally, we used the same procedure to investigate the sources for differences in predicted employment rates and earnings for the pool of immigrants predicted to stay in Canada in each respective year.[30] These results highlight the role played by the outmigration equation coefficients and covariance matrix parameters for unobserved heterogeneity, especially for

[30] See Tables 5 and 6, columns 5–7, for reported differences between LSIC and IMDB in predicted employment and earnings, respectively. Tables with the simulation results are available upon request.

later periods. Overall these results suggest that despite explicitly accounting for measurement errors in attrition, the survey data is unlikely to provide an unbiased picture of the economic performance of immigrants to Canada after landing.

4.4 Additional specifications

We conducted several additional exercises, results of which support two key takeaways from the main analysis: 1) larger non-random attrition unrelated to outmigration in survey data could significantly bias model estimates and predictions, 2) allowing for measurement errors (α_t) in the outmigration indicator only partially reduces this bias.

First, we estimated a restricted specification of the model using LSIC survey data by imposing α_1=0 and α_2=0.[31] The differences in model estimates and predictions using both data sources (LSIC and IMDB data) were found to be even more pronounced under this restriction. For instance, the probabilities of trajectories that were overestimated (underestimated) in our main specification when using LSIC relative to IMDB become even more overestimated (underestimated) under the restricted specification. In particular, the restricted specification overpredicts even to a larger extent immigrants' outmigration. Our main specification suggested that LSIC overpredicts outmigration probability after the first wave by almost 17 percentage points relative to IMDB on average for 16 simulated cases (represented by last two trajectories in Table 4). The restricted specification overpredicts this probability by more than 28 percentage points. The bias in the probability of outmigrating after the second wave (a total of four corresponding trajectories) increases from 0.9 up to 8.1 percentage points. In contrast, the underestimation of the probability of being employed in all three waves (the first trajectory in Table 4) increases from 3.3 to 12.3 percentage points. The mean sum of absolute (squared) deviations between the raw LSIC and IMDB trajectories (reported in Table 7 for the main model) increases from 38.0 (302.0) up to 72.7 (792.8). It follows that allowing for misclassification probabilities reduce the absolute (squared) measure of the overall bias in predicting immigrants' labor market trajectories by 48% (62%).

Second, the main model (allowing for misclassification probabilities) was estimated using the IMDB administrative sample with an alternative measure of attrition that reflects only permanent attrition and thus permanent outmigration. Recall that our administrative sample was

[31] The results from this specification are not reported for the sake of brevity but are available upon request.

constructed to mimic the LSIC survey sample which treats any absence from the panel as final, disregarding whether an immigrant files a return after failing to do so in the wave following the last sample observation. Therefore, the baseline attrition variable in IMDB takes into account even temporary absences from filing a tax declaration and temporary departures. Since administrative data keep track of these individuals, more information is available in IMDB to document presence in the country. Our alternative measure of attrition takes into account whether an individual ever re-appeared in the IMDB dataset after a failing to report a tax declaration. Thus, if an immigrant didn't file taxes in 2003, but did so in 2005 or any subsequent year, the attrition status for that person would be set to zero (in the baseline model it would be set to one for year 2001). Using this measure, which reflects only permanent attrition from the administrative data, the attrition rates in IMDB in $t = 2001$ and $t = 2003$ are 3.1% and 3.0%, respectively (compared to 5.0% and 6.8% using our main coding).[32] We further find that estimates of α_1 and α_2 are smaller using this new coding: 1.0% and 0.5% for 2001 and 2003, respectively.[33] This reflects a smaller measurement error when only permanent outmigration is considered. The differences in predictions of immigrants' labour market trajectories between the LSIC and IMDB in this case increase only marginally, and predictions between two IMDB models using different attrition measures are very close to each other. The mean sum of absolute (squared) deviations between predicted trajectories using two different measures of attrition for the IMDB sample is only 8.0 (9.4).

5 Conclusion

We investigated the economic performances of immigrants to Canada using panel survey (LSIC) and administrative (IMDB) data. We found sizable differences in earnings and employment histories primarily due to the intrinsic nature of sample attrition in each data source. Our administrative data provide more reliable information on outmigration since it tracks immigrants who move within Canada through income tax filing. Similar moves are not to the same extent recorded in the survey data and thus contribute the sample attrition. As with most panel surveys on immigration (see e.g., Bellemare (2007) for Germany, and Cobb-Clark (2001) for Aus-

[32] Aydemir and Robinson (2008) similarly find that shorter absences (less than four years) are more associated with temporary absences from the workforce rather than with definite outmigration, because more than half of immigrants experiencing these shorter spells re-appear in the data.

[33] Estimation results for this attrition measure are not presented in the paper but available upon request.

tralia), the LSIC is plagued with considerable attrition. We conjecture that the results reported in this paper likely generalize more broadly.[34]

We found that the gap in outcomes predicted from both data sources was reduced by incorporating misclassification probabilities when modelling outmigration. As expected, the estimated miscassification probabilities were substantially higher when the model was estimated using survey data, reinforcing the insight that administrative data contains better quality data on outmigration. Our results suggest that incorporating these probabilities is useful when measuring the economic performance of immigrants using survey data. As such, this may represent a second-best alternative to using administrative records which contain more reliable information on outmigration.

With that said, we recognize that administrative records are by no means perfect as they can provide limited information on many social factors (e.g., language fluency) which are important drivers of a successful integration in the host country. A promising research agenda would be to combine survey and administrative data sources to jointly take into account non-random outmigration and the many factors determining the labour market experience of immigrants in the host country.

[34] Additionally,this empirical methodology is not limited to the analysis of the foreign-born population. It could be useful for any longitudinal surveys in countries were emigration rates, including those for natives, are considerable.

References

Agopsowicz, Andrew, Bassirou Gueye, Natalia Kyui, Youngmin Park, Mohanad Salameh, and Ben Tomlin (2017) 'April 2017 annual reassessment of potential output growth in Canada.' *Bank of Canada Staff Analytical Note 2017-5*

Aydemir, Abdurrahman, and Chris Robinson (2008) 'Global labour markets, return, and onward migration.' *Canadian Journal of Economics/Revue canadienne d'économique* 41(4), 1285–1311

Bellemare, Charles (2007) 'A life-cycle model of outmigration and economic assimilation of immigrants in Germany.' *European Economic Review* (51), 553–576

Bijwaard, Govert E (2010) 'Immigrant migration dynamics model for The Netherlands.' *Journal of Population Economics* 23(4), 1213–1247

Bijwaard, Govert E., Christian Schluter, and Jackline Wahba (2014) 'The impact of labor market dynamics on the return migration of immigrants.' *The Review of Economics and Statistics* 96(3), 483–494

Borjas, George J (1985) 'Assimilation, changes in cohort quality, and the earnings of immigrants.' *Journal of labor Economics* 3(4), 463–489

Bryan, Gharad, Shyamal Chowdhury, and Ahmed M. Mobarak (2014) 'Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh.' *Econometrica* 82(5), 1671–1748

Cobb-Clark, Deborah (2001) 'The longitudinal survey of immigrants to australia.' *Australian Economic Review* 34(4), 467–467

Damas de Matos, Ana, and Daniel Parent (2019) 'Canada and high-skill emigration to the united states: Way station or farm system?' *Journal of Labor Economics* 37(S2), S491–S532

Docquier, Frédéric, B. Lindsay Lowell, and Abdeslam Marfouk (2009) 'A gendered assessment of highly skilled emigration.' *Population and Development Review* 35(2), 297–321

Dustmann, Christian, and Albrecht Glitz (2011) 'Chapter 4 - migration and education.' In *Handbook of The Economics of Education*, ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, vol. 4 of *Handbook of the Economics of Education* (Elsevier) pp. 327 – 439

Dustmann, Christian, and Joseph-Simon Görlach (2015) 'Selective out-migration and the estimation of immigrants' earnings profiles.' In 'Handbook of the Economics of International Migration,' vol. 1 (Elsevier) pp. 489–533

Dustmann, Christian, and Yoram Weiss (2007) 'Return migration: Theory and empirical evidence from the UK.' *British Journal of Industrial Relations* 45(2), 236–256

Edo, Anthony, Lionel Ragot, Hillel Rapoport, Sulin Sardoschau, Andreas Steinmayr, and Arthur Sweetman (2020) 'An introduction to the economics of immigration in OECD countries.' *Canadian Journal of Economics/Revue canadienne d'économique* 53(4), 1365–1403

Ferrie, Joseph P, and Timothy J Hatton (2015) 'Two centuries of international migration.' In 'Handbook of the economics of international migration,' vol. 1 (Elsevier) pp. 53–88

Green, David A., and Christopher Worswick (2010) 'Entry earnings of immigrant men in Canada: The roles of labour market entry effects and returns to foreign experience.' *Canadian immigration: Economic evidence for a dynamic policy environment* pp. 77–110

Hausman, J.A., Jason Abrevaya, and F.M. Scott-Morton (1998) 'Misclassification of the dependent variable in a discrete-response setting.' *Journal of Econometrics* 87(2), 239 – 269

Imai, Susumu, Derek Stacey, and Casey Warman (2019) 'From engineer to taxi driver? language proficiency and the occupational skills of immigrants.' *Canadian Journal of Economics/Revue canadienne d'économique* 52(3), 914–953

Kirdar, Murat G (2012) 'Estimating the impact of immigrants on the host country social security system when return migration is an endogenous choice.' *International Economic Review* 53(2), 453–486

Kustec, Stan (2012) 'The role of migrant labour supply in the canadian labour market.' *Citizenship and Immigration Canada*

Lewbel, Arthur (2000) 'Identification of the binary choice model with misclassification.' *Econometric Theory* 16(4), 603–609

Picot, Garnett, and Patrizio Piraino (2013) 'Immigrant earnings growth: Selection bias or real progress.' *Canadian Journal of Economics*

Sweetman, Arthur, and Casey Warman (2013) 'Canada's immigration selection system and labour market outcomes.' *Canadian Public Policy* 39(Supplement 1), S141–S164

Warman, Casey, Arthur Sweetman, and Gustave Goldmann (2015) 'The portability of new immigrants' human capital: Language, education, and occupational skills.' *Canadian Public Policy* 41(Supplement 1), S64–S79

Warman, Casey, Christopher Worswick, and Matthew Webb (2016) 'Immigrant Category of Admission of the Parents and Outcomes of the Children: How far does the Apple Fall?' CReAM Discussion Paper Series 1618, Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London, October

Table 1: Summary Statistics: LSIC and IMDB samples

Variables	LSIC			IMDB		
	2001	2003	2005	2001	2003	2005
Sample Attrition, in %	32.9	18.9	-	5.0	6.8	-
Permanent attrition in administrative data, in %				3.1	3.0	-
Age	35.3	37.1	39.5	35.8	37.9	39.9
Immigration class, in %:						
Principal applicants: non-business class	83.5	84.1	84.5	81.3	81.1	81.0
Principal applicants: business class	4.3	3.2	2.6	4.9	5.0	5.0
Dependents under economic class	12.3	12.7	12.9	13.8	13.9	14.0
Skill level, in %:						
0, A (management & professional occupations)	64.0	61.6	61.5	62.4	62.1	61.7
B, C, D (technical, intermediate, elemental)	23.4	26.4	27.2	22.6	22.7	23.1
Others (e.g., new workers, non-workers)	12.8	11.8	11.0	14.9	15.1	15.2
Education, in %:						
Higher education	84.0	82.4	81.9	78.6	78.6	78.4
Post-secondary education	13.5	15.2	15.9	17.3	17.3	17.5
High school education or below	2.4	2.4	2.3	4.1	4.1	4.1
World area of country of birth, in %:						
Africa, Middle East, South America, Greenland, Atlantic, Pacific and Indian Ocean islands	20.8	23.2	23.9	20.4	20.2	20.0
Asia, Australasia, Pacific	62.4	58.6	57.3	63.6	64.1	64.4
USA, Europe, UK	17.0	18.0	18.4	16.0	15.7	15.7
Province of residence, in %:						
Ontario	57.8	52.8	49.8	57.9	56.8	56.6
Québec	20.9	24.5	25.2	17.7	18.0	17.0
Prairies	8.0	10.3	12.0	8.2	9.3	10.2
British Columbia	13.2	12.4	12.6	16.1	16.0	16.2
Presence in Canada prior to landing as resident, in %	10.0	9.9	10.0	10.9	10.7	10.4
Country of last permanent residency being the same as country of birth, in %	81.5	83.0	82.5	87.7	87.7	87.9
Country of birth characteristics:						
Emigration rate	9.3	9.9	10.4	9.6	9.6	9.6
Mean temperature difference (in $^\circ$C)	21.8	22.2	22.4	21.8	21.8	21.8
Log of Distance to Canada	9.2	9.2	9.1	9.2	9.2	9.2
Unemployment rate	7.2	7.7	7.8	7.1	7.1	7.1
Log of GDP per capita	7.5	7.5	7.5	7.4	7.4	7.4
Log of Population density	4.9	4.8	4.8	4.9	4.9	4.9
Log of non-employment income (when non-zero)	8.4	8.4	8.0	6.9	6.5	6.3
Employed (with annual wages $\geqslant$ 12K, in 2011 CAD), in %	45.9	68.5	78.6	43.0	63.2	72.1
Log of annual wages if employed	10.3	10.6	10.7	10.3	10.5	10.7
Average annual wages if employed (in 2011 CAD)	38729	48907	51961	42900	45700	52300
Number of observations (unweighted)	2250	1503	984	34055	32340	30155
Number of observations (weighted)	35100	23300	15450	34055	32340	30155

Note: For the IMDB, the number of observations is rounded to the nearest five, all percentages are calculated based on rounded-to-five counts, average wages are rounded to the nearest \$100. For the LSIC, the weighted number of observations is rounded to the nearest fifty, and all percentages are calculated using rounded-to-fifty weighted counts. The rounding is done according to Statistics Canada requirements. For the LSIC, the weights of the first wave are used for all reported years.

Source: LSIC and IMDB, authors' calculations.

Table 2a: Parameter estimates of the Attrition/Outmigration Equation

	LSIC				IMDB			
	Coef.	**Std. Err.**	$\Pr > \|z\|$		**Coef.**	**Std. Err.**	$\Pr > \|z\|$	
Individual Characteristics								
Age/10	-0.250	0.132	0.058	*	-0.991	0.190	0.000	***
$(\text{Age}/10)^2$	0.033	0.017	0.052	*	0.130	0.025	0.000	***
Immigration Class								
Principal applicant, non-business class *	-0.050	0.054	0.356		-0.198	0.063	0.002	***
Principal applicant, business class *	0.366	0.052	0.000	***	0.045	0.076	0.554	
Dependents, economic class (*reference*)	-	-	-		-	-	-	
Skill Level								
Skill level: 0 and A *	-0.020	0.058	0.727		0.216	0.071	0.002	***
Skill level: B, C and D *	-0.314	0.063	0.000	***	0.054	0.073	0.462	
Skill level: Others (*reference*)	-	-	-		-	-	-	
Education								
Post-secondary education degree *	-0.023	0.070	0.745		-0.053	0.074	0.476	
Higher education degree *	0.129	0.069	0.060	*	0.017	0.071	0.817	
High school or below (*reference*)	-	-	-		-	-	-	
Province of residence								
Ontario *	0.186	0.029	0.000	***	0.083	0.036	0.023	**
Québec *	-0.223	0.038	0.000	***	-0.023	0.045	0.610	
Prairie provinces *	-0.572	0.061	0.000	***	-0.247	0.058	0.000	***
British Columbia (*reference*)	-	-	-		-	-	-	
Other Individual Characteristics								
Presence in Canada before landing *	0.099	0.032	0.002	***	0.296	0.038	0.000	***
Country of last residence same as of birth *	-0.243	0.027	0.000	***	-0.104	0.036	0.004	***
Number of immigrants in the family	-0.097	0.009	0.000	***	-0.682	0.037	0.000	***
Source Country								
Log of GDP per capita	0.081	0.011	0.000	***	0.126	0.013	0.000	***
Log of population density	-0.076	0.017	0.000	***	0.030	0.020	0.131	
World area 1 *	0.011	0.045	0.802		-0.055	0.051	0.283	
World area 2 *	0.220	0.044	0.000	***	-0.145	0.053	0.007	***
World area 3 (*reference*)	-	-	-		-	-	-	
Emigration rate	-0.001	0.001	0.397		-0.004	0.002	0.004	***
Mean temperature difference	0.006	0.002	0.000	***	0.006	0.002	0.006	***
Log of distance to Canada	0.024	0.060	0.693		-0.018	0.063	0.777	
Unemployment rate	-0.041	0.004	0.000	***	-0.016	0.004	0.000	***
Time								
Year 2003 * (*2001 as the reference*)	-0.383	0.038	0.000	***	0.741	0.050	0.000	***
Dynamic dependence								
Employed in the previous wave *	-0.634	0.075	0.000	***	-0.831	0.062	0.000	***
Measurement errors								
α_1	0.113	0.010	0.000	***	0.016	0.001	0.000	***
α_2	0.116	0.006	0.000	***	0.014	0.002	0.000	***
Constant	-0.441	0.628	0.482		-0.562	0.703	0.424	

Survey based estimates exploit LSIC database. Administrative based estimates exploit IMDB database restricted to the sampling framework mimicking LSIC.
α_1 and α_2 present estimated measurement errors of the outmigration indicators after 2001 and 2003, respectively.
"*" indicates dummy variables.
Source country is defined as the country of birth. World area 1 of the source country includes Africa, Middle East, South America, Greenland, Atlantic, Pacific and Indian Ocean islands; World area 2 includes Asia, Australasia, Pacific; the reference category for the World area variables includes USA, Europe and UK.

Table 2b: Parameter estimates of the Employment Equation

	LSIC				IMDB			
	Coef.	**Std. Err.**	$\Pr > \|z\|$		**Coef.**	**Std. Err.**	$\Pr > \|z\|$	
Individual Characteristics								
Age/10	0.397	0.097	0.000	***	0.407	0.089	0.000	***
$(Age/10)^2$	-0.067	0.013	0.000	***	-0.067	0.011	0.000	***
Immigration Class								
Principal applicant, non-business class *	0.430	0.037	0.000	***	0.235	0.034	0.000	***
Principal applicant, business class *	-0.838	0.048	0.000	***	-0.402	0.036	0.000	***
Dependents, economic class (*reference*)	-	-	-		-	-	-	
Skill Level								
Skill level: 0 and A *	-0.084	0.041	0.039	**	0.012	0.037	0.749	
Skill level: B, C and D *	0.080	0.042	0.054	*	0.300	0.038	0.000	***
Skill level: Others (*reference*)	-	-	-		-	-	-	
Education								
Post-secondary education degree *	-0.358	0.052	0.000	***	0.088	0.036	0.015	**
Higher education degree *	-0.532	0.050	0.000	***	0.068	0.035	0.052	*
High school or below (*reference*)	-	-	-		-	-	-	
Province of residence								
Ontario *	0.232	0.022	0.000	***	0.180	0.018	0.000	***
Québec *	-0.714	0.027	0.000	***	-0.412	0.024	0.000	***
Prairie provinces *	0.398	0.031	0.000	***	0.406	0.026	0.000	***
British Columbia (*reference*)	-	-	-		-	-	-	
Other Individual Characteristics								
Presence in Canada before landing *	0.806	0.029	0.000	***	0.649	0.023	0.000	***
Lof of non-employment income	-0.051	0.002	0.000	***	-0.047	0.001	0.000	***
Source Country								
Log of GDP per capita	-0.060	0.008	0.000	***	-0.016	0.007	0.020	**
Log of population density	0.058	0.009	0.000	***	0.088	0.008	0.000	***
World area 1 *	-0.469	0.025	0.000	***	-0.669	0.024	0.000	***
World area 2 *	-0.818	0.028	0.000	***	-0.778	0.026	0.000	***
World area 3 (*reference*)	-	-	-		-	-	-	
Time								
Year 2003 *	0.533	0.022	0.000	***	0.546	0.018	0.000	***
Year 2005 *	0.873	0.031	0.000	***	0.803	0.025	0.000	***
Year 2001 (*reference*)	-	-	-		-	-	-	
Dynamic dependence								
Employed in the previous wave *	0.703	0.027	0.000	***	0.604	0.022	0.000	***
Constant	0.250	0.209	0.233		-0.762	0.189	0.000	***

Survey based estimates exploit LSIC database. Administrative based estimates exploit IMDB database restricted to the sampling framework mimicking LSIC.
"*" indicates dummy variables. Source country is defined as the country of birth. World area 1 of the source country includes Africa, Middle East, South America, Greenland, Atlantic, Pacific and Indian Ocean islands; World area 2 includes Asia, Australasia, Pacific; the reference category for the World area variables includes USA, Europe and UK.

Table 2c: Parameter Estimates of the Earnings Equation

	LSIC				IMDB			
	Coef.	**Std. Err.**	$\Pr > \|z\|$		**Coef.**	**Std. Err.**	$\Pr > \|z\|$	
Individual Characteristics								
Age/10	0.371	0.040	0.000	***	0.302	0.043	0.000	***
$(\text{Age}/10)^2$	-0.052	0.005	0.000	***	-0.042	0.005	0.000	***
Immigration Class								
Principal applicant, non-business class *	0.097	0.017	0.000	***	0.116	0.018	0.000	***
Principal applicant, business class *	-0.178	0.027	0.000	***	-0.133	0.022	0.000	***
Dependents, economic class (*reference*)	-	-	-		-	-	-	
Skill Level								
Skill level: 0 and A *	0.138	0.020	0.000	***	0.137	0.020	0.000	***
Skill level: B, C and D *	0.048	0.021	0.020	**	0.067	0.020	0.001	***
Skill level: Others (*reference*)	-	-	-		-	-	-	
Education								
Post-secondary education degree *	0.095	0.024	0.000	***	0.092	0.019	0.000	***
Higher education degree *	0.141	0.023	0.000	***	0.149	0.019	0.000	***
High school or below (*reference*)	-	-	-		-	-	-	
Province of residence								
Ontario *	0.157	0.010	0.000	***	0.122	0.009	0.000	***
Québec *	-0.126	0.012	0.000	***	-0.165	0.012	0.000	***
Prairie provinces *	0.230	0.012	0.000	***	0.182	0.012	0.000	***
British Columbia (*reference*)	-	-	-		-	-	-	
Other Individual Characteristics								
Presence in Canada before landing *	0.384	0.011	0.000	***	0.470	0.010	0.000	***
Source Country								
Log of GDP per capita	0.098	0.004	0.000	***	0.075	0.004	0.000	***
Log of population density	0.002	0.004	0.616		0.015	0.004	0.000	***
World area 1 *	-0.097	0.011	0.000	***	-0.197	0.012	0.000	***
World area 2 *	-0.176	0.012	0.000	***	-0.275	0.013	0.000	***
World area 3 (*reference*)	-	-	-		-	-	-	
Time								
Year 2003 *	0.319	0.007	0.000	***	0.277	0.007	0.000	***
Year 2005 *	0.439	0.010	0.000	***	0.443	0.009	0.000	***
Year 2001 (*reference*)	-	-	-		-	-	-	
Dynamic dependence								
Employed in the previous wave *	0.054	0.007	0.000	***	0.108	0.007	0.000	***
Constant	8.447	0.092	0.000	***	8.675	0.093	0.000	***

Survey based estimates exploit LSIC database. Administrative based estimates exploit IMDB database restricted to the sampling framework mimicking LSIC.
"*" indicates dummy variables.
Source country is defined as the country of birth. World area 1 of the source country includes Africa, Middle East, South America, Greenland, Atlantic, Pacific and Indian Ocean islands; World area 2 includes Asia, Australasia, Pacific; the reference category for the World area variables includes USA, Europe and UK.

Table 2d: Estimated Covariance Matrices

	LSIC				IMDB			
	Coef.	**Std. Err.**	$\Pr > \|z\|$		**Coef.**	**Std. Err.**	$\Pr > \|z\|$	
Stochastic Terms								
$\sigma^2_{\varepsilon_w}$	0.086	0.001	0.000	***	0.134	0.001	0.000	***
$\rho_{\varepsilon_r\varepsilon_e}$	-0.399	0.021	0.000	***	-0.766	0.036	0.000	***
$\rho_{\varepsilon_r\varepsilon_w}$	-0.087	0.038	0.021	**	-0.394	0.028	0.000	***
$\rho_{\varepsilon_e\varepsilon_w}$	-0.175	0.027	0.000	***	0.014	0.038	0.720	
Unobserved Heterogeneity Parameters								
$\sigma^2_{\eta_r}$	0.094	0.036	0.009	***	0.286	0.058	0.000	***
$\sigma^2_{\eta_e}$	0.522	0.032	0.000	***	0.524	0.025	0.000	***
$\sigma^2_{\eta_w}$	0.222	0.004	0.000	***	0.225	0.004	0.000	***
$\rho_{\eta_r\eta_e}$	-0.513	0.038	0.000	***	0.766	0.043	0.000	***
$\rho_{\eta_r\eta_w}$	-0.671	0.015	0.000	***	0.546	0.027	0.000	***
$\rho_{\eta_e\eta_w}$	0.870	0.013	0.000	***	0.839	0.013	0.000	***

Table 3: Predicted Distribution of Labour Market Transitions Using Administrative Data (IMDB)

Prov-Age-Educ	**(e,e,e)**	**(e,e,n)**	**(e,e,.)**	**(e,n,e)**	**(e,n,n)**	**(e,n,.)**	**(n,e,e)**	**(n,e,n)**	**(n,e,.)**	**(n,n,e)**	**(n,n,n)**	**(n,n,.)**	**(e,.,.)**	**(n,.,.)**
ON-25-PSE	43.2%	2.6%	1.1%	3.0%	1.4%	0.6%	22.1%	3.2%	0.7%	8.9%	7.4%	2.4%	0.8%	2.6%
QC-25-PSE	27.9%	2.8%	0.9%	3.1%	2.1%	0.7%	21.3%	4.8%	0.9%	12.1%	15.2%	4.3%	0.6%	3.4%
BC-25-PSE	33.1%	2.8%	0.7%	3.2%	1.9%	0.5%	22.7%	4.4%	0.6%	11.8%	12.7%	2.9%	0.4%	2.3%
PR-25-PSE	56.9%	2.0%	0.8%	2.6%	0.9%	0.3%	20.9%	2.0%	0.4%	6.5%	3.7%	1.0%	0.6%	1.3%
ON-35-PSE	43.4%	2.8%	0.8%	3.2%	1.5%	0.4%	22.4%	3.4%	0.5%	9.3%	8.1%	1.9%	0.5%	1.7%
QC-35-PSE	28.1%	3.1%	0.6%	3.2%	2.3%	0.6%	21.4%	5.2%	0.6%	12.5%	16.4%	3.5%	0.3%	2.3%
BC-35-PSE	33.1%	3.0%	0.5%	3.4%	2.1%	0.4%	22.7%	4.7%	0.4%	12.1%	13.6%	2.3%	0.2%	1.5%
PR-35-PSE	57.2%	2.2%	0.5%	2.8%	1.0%	0.2%	21.0%	2.1%	0.3%	6.7%	4.1%	0.8%	0.4%	0.8%
ON-25-HE	42.3%	2.6%	1.2%	2.9%	1.4%	0.6%	21.9%	3.2%	0.8%	8.8%	7.6%	2.7%	0.9%	3.0%
QC-25-HE	27.2%	2.9%	1.0%	3.0%	2.1%	0.8%	20.8%	4.9%	1.0%	11.8%	15.3%	4.8%	0.6%	3.9%
BC-25-HE	32.3%	2.8%	0.8%	3.2%	1.9%	0.6%	22.3%	4.5%	0.7%	11.6%	12.9%	3.2%	0.5%	2.7%
PR-25-HE	56.0%	2.0%	0.9%	2.7%	0.9%	0.3%	20.9%	2.0%	0.5%	6.5%	3.9%	1.1%	0.7%	1.5%
ON-35-HE	42.7%	2.8%	0.9%	3.1%	1.5%	0.5%	22.1%	3.5%	0.6%	9.3%	8.2%	2.2%	0.6%	2.0%
QC-35-HE	27.3%	3.1%	0.7%	3.2%	2.2%	0.6%	21.1%	5.2%	0.7%	12.2%	16.7%	3.9%	0.4%	2.6%
BC-35-HE	32.4%	3.1%	0.5%	3.3%	2.1%	0.5%	22.4%	4.7%	0.5%	12.0%	13.9%	2.5%	0.3%	1.7%
PR-35-HE	56.2%	2.2%	0.6%	2.8%	1.0%	0.3%	21.1%	2.2%	0.3%	6.8%	4.2%	0.9%	0.4%	1.0%

The columns present the predicted shares of labour market transitions over the three sampling years (2001,2003,2005), where **e** denotes employment, **n** denotes unemployment, and · captures outmigration. The rows indicate the conditioning province (ON=Ontario,QC=Québec,BC=British Columbia,PR=Prairie provinces), age and education levels (PSE=post-secondary education, HE=higher education).

Table 4: Percentage Point Differences in Predicted Labour Market Transitions Between the LSIC and IMDB models

Prov-Age-Educ	**(e,e,e)**	**(e,e,n)**	**(e,e,.)**	**(e,n,e)**	**(e,n,n)**	**(e,n,.)**	**(n,e,e)**	**(n,e,n)**	**(n,e,.)**	**(n,n,e)**	**(n,n,n)**	**(n,n,.)**	**(e,.,.)**	**(n,.,.)**
ON-25-PSE	-1.5%	-1.2%	0.6%	-1.0%	-0.7%	0.0%	-9.1%	-2.1%	0.9%	-4.7%	-5.1%	0.1%	9.6%	14.1%
QC-25-PSE	-1.3%	-0.6%	-0.7%	0.0%	-0.2%	-0.5%	-1.1%	-1.3%	-0.3%	0.9%	-1.4%	-1.5%	1.1%	6.9%
BC-25-PSE	-2.2%	-1.0%	0.2%	-0.7%	-0.6%	0.0%	-6.8%	-2.3%	0.9%	-4.1%	-6.3%	1.1%	5.4%	16.2%
PR-25-PSE	2.9%	-0.8%	-0.4%	-0.5%	-0.3%	-0.2%	-3.6%	-1.0%	0.1%	-1.6%	-1.6%	-0.1%	2.9%	4.2%
ON-35-PSE	-1.4%	-1.2%	0.8%	-1.1%	-0.8%	0.1%	-9.0%	-2.2%	1.1%	-4.9%	-5.4%	0.6%	9.2%	14.2%
QC-35-PSE	-1.9%	-0.7%	-0.4%	-0.1%	-0.3%	-0.4%	-1.5%	-1.4%	-0.1%	0.5%	-1.4%	-0.8%	1.1%	7.4%
BC-35-PSE	-2.3%	-1.0%	0.4%	-0.7%	-0.7%	0.1%	-6.7%	-2.3%	0.9%	-4.2%	-6.5%	1.7%	5.1%	16.2%
PR-35-PSE	2.5%	-0.8%	-0.2%	-0.6%	-0.4%	-0.1%	-3.5%	-1.0%	0.2%	-1.6%	-1.7%	0.0%	2.8%	4.3%
ON-25-HE	-7.3%	-1.1%	0.7%	-0.9%	-0.6%	0.0%	-9.6%	-2.0%	1.2%	-4.4%	-4.7%	0.7%	10.2%	18.0%
QC-25-HE	-5.8%	-0.7%	-0.8%	-0.2%	-0.1%	-0.6%	-2.5%	-1.1%	-0.3%	1.2%	1.1%	-1.0%	1.1%	9.5%
BC-25-HE	-7.3%	-1.0%	0.2%	-0.8%	-0.5%	0.0%	-8.0%	-2.2%	1.0%	-4.0%	-5.4%	1.9%	5.6%	20.4%
PR-25-HE	-2.7%	-0.6%	-0.4%	-0.3%	-0.2%	-0.1%	-2.9%	-0.8%	0.2%	-0.7%	-0.9%	0.1%	3.4%	5.9%
ON-35-HE	-7.4%	-1.1%	0.9%	-1.0%	-0.6%	0.1%	-9.5%	-2.1%	1.3%	-4.6%	-5.0%	1.2%	9.8%	18.1%
QC-35-HE	-6.3%	-0.8%	-0.5%	-0.3%	-0.1%	-0.4%	-3.2%	-1.2%	-0.1%	0.9%	1.0%	-0.1%	1.2%	9.9%
BC-35-HE	-7.4%	-1.0%	0.4%	-0.8%	-0.5%	0.2%	-8.1%	-2.2%	1.1%	-4.2%	-5.6%	2.6%	5.3%	20.3%
PR-35-HE	-3.0%	-0.6%	-0.1%	-0.4%	-0.2%	-0.1%	-3.0%	-0.8%	0.3%	-0.8%	-0.9%	0.3%	3.3%	6.0%

The columns present the percentage point differences in predicted shares of labour market transitions over the three sampling years (2001,2003,2005), where **e** denotes employment, **n** denotes unemployment, and · captures outmigration. The rows indicate the conditioning province (ON=Ontario,QC=Québec,BC=British Columbia,PR=Prairie provinces), age and education levels (PSE=post-secondary education, HE=higher education). Predicted negative/positive differences exceeding 2 percentage points are highlighted in pink and yellow, respectively.

Table 5: Predicted Employment Probabilities in the IMDB Model, and Differences Between the LSIC and IMDB Models, in percentage points.

	Predicted employment probabilities for those who are predicted to be present in Canada						**Predicted employment probabilities for those predicted to be present in Canada per year and:** *predicted to outmigrate before the following period*				*predicted to stay in the following period*			
	IMDB, %			LSIC vs IMDB, pp			IMDB		LSIC vs IMDB, pp		IMDB		LSIC vs IMDB, pp	
	2001	**2003**	**2005**	**2001**	**2003**	**2005**	**2001**	**2003**	**2001**	**2003**	**2001**	**2003**	**2001**	**2003**
ON-25-PSE	53%	75%	84%	5.8	7.6	7.6	24%	38%	14.9	14.1	54%	77%	12.2	8.7
QC-25-PSE	38%	61%	72%	-2.3	-0.4	2.5	14%	26%	-0.6	-4.8	39%	64%	-0.2	-1.3
BC-25-PSE	43%	66%	76%	1.2	4.3	6.6	16%	28%	8.3	7.6	43%	68%	6.9	5.9
PR-25-PSE	64%	85%	91%	3.6	3.6	3.5	32%	49%	7.3	0.5	65%	86%	5.8	3.5
ON-35-PSE	53%	75%	83%	5.5	7.4	7.6	23%	36%	15.3	14.6	53%	77%	11.8	8.8
QC-35-PSE	38%	61%	71%	-2.7	-1.1	2.1	13%	24%	0.2	-4.4	39%	63%	-0.6	-1.5
BC-35-PSE	43%	66%	75%	0.8	3.8	6.3	14%	26%	8.9	7.6	43%	67%	6.4	5.8
PR-35-PSE	64%	84%	90%	3.4	3.5	3.5	30%	45%	8.7	2.3	65%	85%	5.6	3.5
ON-25-HE	52%	75%	84%	1.0	4.4	5.7	24%	39%	11.1	10.8	53%	77%	8.4	6.2
QC-25-HE	38%	60%	71%	-6.9	-5.5	-1.9	14%	26%	-2.2	-7.8	39%	63%	-4.6	-6.2
BC-25-HE	42%	66%	76%	-3.7	-0.3	3.3	16%	28%	5.0	3.8	43%	68%	2.6	2.1
PR-25-HE	64%	84%	91%	-1.0	0.8	2.0	32%	49%	3.6	-2.8	64%	85%	1.9	1.0
ON-35-HE	52%	74%	83%	0.6	3.9	5.6	22%	36%	11.7	10.8	53%	76%	8.0	6.1
QC-35-HE	38%	60%	70%	-7.1	-6.4	-2.4	13%	24%	-1.7	-7.0	38%	62%	-4.7	-6.5
BC-35-HE	42%	65%	75%	-4.0	-0.9	2.8	14%	26%	6.2	4.6	43%	67%	2.3	1.8
PR-35-HE	64%	84%	90%	-1.2	0.5	1.8	30%	46%	4.8	-1.3	64%	85%	1.7	0.9
Average	49%	71%	80%	-0.4	1.6	3.5	21%	34%	6.3	3.0	50%	73%	4.0	2.4

Rows present predictions by province (ON=Ontario,QC=Québec,BC=British Columbia,PR=Prairie provinces), age and education levels (PSE=post-secondary education, HE=higher education). Predicted negative/positive differences exceeding 2 percentage points are highlighted in pink and yellow, respectively.

Table 6: Predicted Earnings in the IMDB Model and Differences Between the LSIC and the IMDB Models, in percent

	Predicted earnings for those who are predicted to be present in Canada and employed in the reported year						**Predicted earnings for those predicted to be present and employed in the reported year and:** *predicted to outmigrate before the following period*				*predicted to stay for the following period*			
	IMDB, CAD			LSIC vs IMDB, %			IMDB		LSIC vs IMDB		IMDB		LSIC vs IMDB	
	2001	**2003**	**2005**	**2001**	**2003**	**2005**	**2001**	**2003**	**2001**	**2003**	**2001**	**2003**	**2001**	**2003**
ON-25-PSE	31342	41961	49365	1.3%	8.5%	5.7%	35671	40333	-20.1%	-11.7%	31275	42003	3.8%	9.8%
QC-25-PSE	28128	36991	43335	4.2%	10.0%	5.5%	32660	35963	-22.1%	-13.6%	28061	37022	5.1%	10.3%
BC-25-PSE	27765	36748	43199	-0.6%	6.1%	2.7%	31651	35857	-21.4%	-15.0%	27725	36766	1.0%	7.2%
PR-25-PSE	35919	48927	57879	2.8%	7.0%	2.8%	40848	46163	-22.3%	-19.4%	35871	48967	3.8%	7.2%
ON-35-PSE	32957	43534	50496	2.5%	9.1%	5.7%	38351	41938	-21.4%	-11.1%	32905	43563	4.9%	10.3%
QC-35-PSE	29508	38379	44371	5.8%	10.8%	5.7%	34655	37098	-21.2%	-12.7%	29463	38407	6.6%	11.1%
BC-35-PSE	29085	38058	44171	0.7%	6.8%	2.6%	33614	36496	-21.9%	-12.6%	29059	38081	2.3%	7.7%
PR-35-PSE	37577	50550	58900	4.7%	7.9%	3.2%	42282	48601	-20.7%	-20.1%	37551	50569	5.5%	8.2%
ON-25-HE	33369	44549	52412	1.3%	9.0%	6.5%	38281	42586	-19.8%	-9.6%	33281	44607	4.0%	10.7%
QC-25-HE	29877	39231	45903	4.8%	11.0%	6.6%	34070	37733	-17.9%	-10.1%	29807	39284	5.7%	11.3%
BC-25-HE	29508	38946	45753	-0.4%	6.9%	3.9%	34106	37825	-22.3%	-11.8%	29452	38973	1.6%	8.2%
PR-25-HE	38026	51724	61167	3.3%	7.3%	3.1%	42254	48859	-19.3%	-17.2%	37979	51773	4.4%	7.7%
ON-35-HE	34951	46137	53458	2.9%	9.7%	6.5%	39556	44149	-18.0%	-8.9%	34901	46181	5.5%	11.2%
QC-35-HE	31359	40664	47021	6.4%	12.0%	6.6%	36842	39568	-20.2%	-11.3%	31303	40691	7.3%	12.4%
BC-35-HE	30911	40381	46734	1.1%	7.6%	4.0%	36194	39260	-21.9%	-11.3%	30875	40400	2.9%	8.8%
PR-35-HE	39886	53561	62364	4.8%	8.0%	3.3%	45230	51325	-19.6%	-17.5%	39851	53587	5.7%	8.4%
Average	32511	43146	50408	2.9%	8.6%	4.6%	37266	41485	-20.6%	-13.6%	32460	43180	4.4%	9.3%

Rows present predictions by province (ON=Ontario,QC=Québec,BC=British Columbia,PR=Prairie provinces), Age and education levels (PSE=post-secondary education, HE=higher education). Predicted negative/positive differences exceeding 5 percent are highlighted in pink and yellow, respectively.

Table 7: Differences between LSIC and IMDB Trajectories Due to Parameter Estimates

Estimated parameters used in simulations	**Sum of Absolute Deviations**				**Sum of Squared Deviations**			
	Mean	**Min**	**Max**	**ΔMean, in %**	**Mean**	**Min**	**Max**	**ΔMean, in %**
LSIC ($\gamma^L, \delta^L, \beta^L, \Sigma_\varepsilon^L, \Sigma_\eta^L$) vs IMDB ($\gamma^I, \delta^I, \beta^I, \Sigma_\varepsilon^I, \Sigma_\eta^I$)	38.0	17.7	62.8	100.0%	302.0	52.7	625.4	100.0%
1 element from IMDB estimates (other four - from LSIC):								
Outmigration coefficients (γ^I)	11.9	5.2	19.1	-68.7%	28.8	2.6	89.8	-90.4%
Employment coefficients (δ^I)	36.8	14.5	63.1	-3.1%	303.0	39.5	656.9	0.3%
Wage coefficients (β^I)	38.0	17.8	63.0	0.1%	302.1	52.2	629.7	0.0%
Covariance matrix Σ_ε^I	38.9	17.1	61.9	2.5%	382.7	77.7	764.9	26.7%
Covariance matrix Σ_η^I	41.9	18.1	64.4	10.2%	350.1	58.8	731.7	15.9%
2 elements from IMDB estimates (other three - from LSIC):								
γ^I and δ^I	5.7	2.9	8.4	-85.0%	3.8	0.8	7.7	-98.8%
γ^I and Σ_η^I	11.7	5.4	20.1	-69.2%	27.3	2.9	87.0	-91.0%
γ^I and β^I	11.9	5.3	19.2	-68.6%	29.0	2.5	91.8	-90.4%
γ^I and Σ_ε^I	13.1	5.8	21.7	-65.6%	33.5	3.8	109.5	-88.9%
δ^I and β^I	36.8	14.5	62.9	-3.2%	301.9	38.8	650.7	0.0%
δ^I and Σ_ε^I	37.1	14.9	62.0	-2.3%	380.2	59.3	782.6	25.9%
β^I and Σ_ε^I	38.9	17.2	61.9	2.5%	381.6	78.2	762.7	26.4%
δ^I and Σ_η^I	39.8	18.3	64.3	4.7%	354.5	78.3	774.1	17.4%
Σ_ε^I and Σ_η^I	40.9	19.5	63.5	7.5%	353.1	64.6	702.2	16.9%
β^I and Σ_η^I	41.8	18.0	64.4	10.1%	350.4	58.2	732.8	16.0%
3 elements from IMDB estimates (other two - from LSIC):								
γ^I, δ^I, and Σ_η^I	5.0	2.4	7.7	-86.8%	3.0	0.7	6.2	-99.0%
γ^I, δ^I, and β^I	5.7	2.9	8.6	-85.1%	3.7	0.9	8.0	-98.8%
γ^I, δ^I, and Σ_ε^I	7.5	4.0	11.1	-80.3%	8.0	2.1	16.4	-97.3%
γ^I, Σ_ε^I, and Σ_η^I	9.3	2.9	17.3	-75.7%	24.7	1.4	70.6	-91.8%
γ^I, β^I, and Σ_η^I	11.7	5.4	20.1	-69.2%	27.3	2.9	87.0	-91.0%
γ^I, β^I, and Σ_ε^I	13.1	5.8	21.7	-65.6%	33.5	3.8	109.5	-88.9%
δ^I, β^I, and Σ_ε^I	37.1	15.0	61.7	-2.4%	379.1	59.8	776.5	25.5%
δ^I, Σ_ε^I, and Σ_η^I	39.1	17.7	63.7	3.0%	356.0	67.9	745.3	17.9%
δ^I, β^I, and Σ_η^I	39.8	18.3	64.5	4.7%	355.0	79.3	776.5	17.5%
β^I, Σ_ε^I, and Σ_η^I	40.9	19.5	63.6	7.5%	353.3	63.8	706.6	17.0%
4 elements from IMDB estimates (remaining one - from LSIC):								
γ^I, δ^I, Σ_ε^I, and Σ_η^I	0.3	0.2	0.5	-99.1%	0.02	0.01	0.05	-100.0%
γ^I, δ^I, β^I, and Σ_η^I	5.0	2.6	7.7	-86.7%	3.0	0.8	6.0	-99.0%
γ^I, δ^I, β^I, and Σ_ε^I	7.5	3.9	11.0	-80.3%	8.0	2.0	16.4	-97.4%
γ^I, β^I, Σ_ε^I, and Σ_η^I	9.3	2.8	17.3	-75.6%	24.9	1.3	71.2	-91.8%
δ^I, β^I, Σ_ε^I, and Σ_η^I	39.2	17.8	63.8	3.1%	356.3	68.4	745.1	18.0%

Source: Authors' calculations.

A Appendix: Details on source countries' variables

In the estimation, we use the following variables that describe characteristics of immigrants' source countries (countries of birth).

- **Distance to Canada**. For measuring the distance of immigrants' source country (country of birth) to Canada, we use the GeoDist data from the CEPII research center (Mayer and Zignago (2011) describe in detail the construction of the dataset).[35] Specifically, we use the CEPII distance measure constructed based on bilateral distances between the biggest cities of those two countries weighted by the share of cities in the population of those countries. This indicator varies across countries, from around 2000 up to more than 15000, with both mean and median being around 9000.

- **GDP per capita**. We use the United Nations dataset "National Accounts Main Aggregates Database" to extract GDP per capita in the country of birth of an immigrant.[36] In particular, we calculate the measure of GDP per capita expressed in constant 2005 US dollars, in an immigrant's country of birth in a year prior to the immigrant's landing in Canada. Therefore, this variable aims at capturing economic conditions in the country of birth prior to the immigrant's move to Canada. Over 1999-2000 this indicator varied across countries from $120 up to $120K, with the mean value being around $11K and the median around $3K.

- **Temperature difference**. We use data on mean temperatures by countries derived from the Climate Change Knowledge Portal of the World Bank.[37] In particular we use the temperature averages for the period 1961–1999. The variable used in the estimations is the difference between mean temperatures in an immigrant's country of birth and that of Canada. Across all countries the indicator of the difference in mean temperatures relative to Canada varies from -9°C up to +35°C, with the mean being +25°C, and the median being +29°C.

- **Population Density**. We use population density in a source country, defined as mid-year population divided by land area in square kilometers, derived from the "World Devel-

[35] Thierry Mayer and Soledad Zignago (2011), "Notes on CEPII's distances measures: The GeoDist database", CEPII Working Paper, No 2011-25. For the center information, see http://www.cepii.fr.

[36] http://unstats.un.org/unsd/snaama/resQuery.asp

[37] http://data.worldbank.org/data-catalog/cckp_historical_data

opment Indicators" database of the World Bank.[38] We merge the indicator of population density in the country of birth corresponding to the year prior to the landing year of an immigrant. Across all countries, this indicator in 1999-2000 varied from 0.14 up to more than 21000, with the mean around 330 and median around 62 persons per sq.km.

- **Unemployment Rate**. We use the unemployment rate indicator estimated by the International Labour Organisation (data are available in the "World Development Indicators" database of the World Bank). In order to remove the effect of the business cycle on the unemployment rate over time, we average unemployment rates over 1991–2011; thus, this variable captures level differences across countries in the unemployment measure. The ILO's estimates of the unemployment rate, from the ILO's Key Indicators of the Labour Market database, are based on household labour force surveys, including both reported and imputed data. Importantly, the ILO estimates account for differences that may alter comparability of this indicator across countries, including differences in definition, data source, coverage, period, and methodologies of data collection. Across all countries, the constructed unemployment indicator varies from 0.6% to 33.3%, with the mean being 8.9%, and the median being 7.7%.

- **Emigration Rate**. In order to capture differences across source countries in the propensity of populations to emigrate, we use the emigration rate of the tertiary educated population, defined as the fraction of emigrants in the population with tertiary education. Data come from the "World Development Indicators" database of the World Bank, and are based on Docquier, Lowell and Marfouk (2009)[39] In particular, this indicator is calculated as the stock of emigrants aged 25 and older, residing in an OECD country other than that in which they were born, with at least one year of tertiary education as a percentage of the 25+ population with tertiary education in a respective country of birth. We use this measure calculated for the year 2000. Across all countries, the emigration rate in 2000 varied from 0.4% to 89.2%, with the mean at 19.8% and the median at 10.9%.

[38] http://data.worldbank.org/data-catalog/world-development-indicators

[39] Frederic Docquier, B. Lindsay Lowell, and Abdeslam Marfouk (2009), "A Gendered Assessment of Highly Skilled Emigration", Population and Development Review, Volume 35, Issue 2, June 2009, Pages 297–321.

B Appendix: Summary statistics tables

Table B1: Sources of information on variables in LSIC and IMDB databases

Variables	Source of information	
	LSIC	**IMDB**
Dependent variables:		
Sample attrition	Derived from sample years	Derived from sample years
Permanent attrition	N/A	Derived from all available years
Employment: working with annual earnings above or equal to 12K (in 2011 CAD) Wages: annual wage earnings (if employed)	Self-reported	Income-tax declarations T1 (includes T4 earnings and other employment income)
Explanatory variables:		
Age	Self-reported (all-waves)	Inferred from the year of birth (FOSS records)
Province of residence	Self-reported	Income-tax declarations T1
Immigration class	FOSS records	
Skill level	FOSS records	
Education	Self-reported (1st wave)	FOSS records
World area of country of birth	FOSS records	
Country of last permanent residency being the same as country of birth	FOSS records	
Presence in Canada prior to landing as permanent resident	Self-reported (1st wave): presence on a non-tourist status	Income-tax declarations T1: filed declarations before landing
Number of other immigrants in a family	Self-reported	Income-tax declarations T1
Log of non-employment income	Self-reported	Income-tax declarations T1
Country of birth characteristics:	Country of birth is derived from FOSS records	
Emigration rate	"World Development Indicators" database of the World Bank	
Log of Population density	"World Development Indicators" database of the World Bank	
Unemployment rate	"World Development Indicators" database of the World Bank	
Mean temperature difference (in $^{\circ}C$)	Climate Change Knowledge Portal of the World Bank	
Log of Distance to Canada	Databse of CEPII research center	
Log of GDP per capita	United Nations "National Accounts Main Aggregates Database"	

C Appendix: Maximum Likelihood function

For simplicity, denote $Y_i = \{e_{it}, w_{it}, r_{it}\}_{t=1}^3$ as the set of dependent variables, $X_i = \{X_{it}^e, X_{it}^w, X_{it}^o\}_{t=1}^3$ as the set of all exogenous observable characteristics, and $\theta = \{\alpha_t, \delta, \beta, \gamma, \Sigma_\varepsilon\}_{t=1}^3$ as the set of model parameters to be estimated. Note that for $t = 3$, r_{it}, X_{it}^o and α_t are excluded from from the model since the attrition status in the last period is not observed. We first write the likelihood function conditional on all time-invariant unobserved heterogeneity parameters $\eta_i = (\eta_i^1, \eta_i^2, \eta_i^3)$. The unconditional likelihood function is next obtained by integrating them out.

The joint density of the employment status, earnings and attrition for the next period, conditional on η_i, is given by:

$$
\begin{aligned}
& f^c(Y_i|X_i, \theta, \eta_i) = f_1^c \cdot f_2^c \cdot f_3^c, \text{ where for } t = \{1, 2\}: \qquad \text{(C.1)} \\
& f_t^c = \Big[(P(e_{it} = 0, r_{it} = 1))^{1-e_{it}} \cdot (P(e_{it} = 1, w_{it}, r_{it} = 1))^{e_{it}}\Big]^{r_{it}} \cdot \\
& \qquad\qquad \times \Big[(P(e_{it} = 0, r_{it} = 0))^{1-e_{it}} \cdot (P(e_{it} = 1, w_{it}, r_{it} = 0))^{e_{it}}\Big]^{1-r_{it}}; \\
& = \Big[(P(e_{it}^* < 0, r_{it} = 1))^{1-e_{it}} \cdot (P(e_{it}^* \geqslant 0, w_{it}, r_{it} = 1))^{e_{it}}\Big]^{r_{it}} \cdot \\
& \qquad\qquad \times \Big[(P(e_{it}^* < 0, r_{it} = 0))^{1-e_{it}} \cdot (P(e_{it}^* \geqslant 0, w_{it}, r_{it} = 0))^{e_{it}}\Big]^{1-r_{it}}; \\
& \text{and for } t = 3: \\
& f_3^c = [P(e_{it} = 0)]^{1-e_{it}} \cdot [P(e_{it} = 1, w_{it})]^{e_{it}} = [P(e_{it}^* < 0)]^{1-e_{it}} \cdot [P(e_{it}^* \geqslant 0, w_{it})]^{e_{it}}.
\end{aligned}
$$

The conditioning on $\{X_i, \theta, \eta_i\}$ is omitted for simplicity, but is assumed in all probabilistic expressions. The above probabilities can be calculated as follows:

$$
\begin{aligned}
P(e_{it}^* < 0, r_{it} = 1) &= \alpha_t P(e_{it}^* < 0) + (1-\alpha_t) P(e_{it}^* < 0, r_{it}^{o*} \geqslant 0) \; ; \qquad \text{(C.2)} \\
P(e_{it}^* < 0, r_{it} = 0) &= (1-\alpha_t) P(e_{it}^* < 0, r_{it}^{o*} < 0) \; ; \\
P(e_{it}^* < 0) &= \Phi(-X_{it}^{e'}\delta - \eta_i^2) \; ; \\
P(e_{it}^* < 0, r_{it}^{o*} \geqslant 0) &= G(-X_{it}^{e'}\delta - \eta_i^2; \quad X_{it}^{o'}\gamma + \eta_i^1; \; -\rho_{12}) \; ; \\
P(e_{it}^* < 0, r_{it}^{o*} < 0) &= G(-X_{it}^{e'}\delta - \eta_i^2; \; -X_{it}^{o'}\gamma - \eta_i^1; \quad \rho_{12}) \; ; \\
P(e_{it}^* \geqslant 0, w_{it}, r_{it} = 1) &= \alpha_t P(e_{it}^* \geqslant 0, w_{it}) + (1-\alpha_t) P(e_{it}^* \geqslant 0, w_{it}, r_{it}^{o*} \geqslant 0) \; ; \\
P(e_{it}^* \geqslant 0, w_{it}, r_{it} = 0) &= (1-\alpha_t) P(e_{it}^* \geqslant 0, w_{it}, r_{it}^{o*} < 0) \; ; \\
P(e_{it}^* \geqslant 0, w_{it}) &= \frac{1}{\sigma_{\varepsilon_3}} \phi\left(\frac{w_{it} - X_{it}^{w'}\beta - \eta_i^3}{\sigma_{\varepsilon_3}} \right) \Phi\left(\frac{X_{it}^{e'}\delta + \eta_i^2 + \frac{\rho_{23}}{\sigma_{\varepsilon_3}}(w_{it} - X_{it}^{w'}\beta - \eta_i^3)}{\sqrt{1-\rho_{23}^2}} \right) \; ; \\
P(e_{it}^* \geqslant 0, w_{it}, r_{it}^{o*} \geqslant 0) &= \frac{1}{\sigma_{\varepsilon_3}} \phi\left(\frac{w_{it} - X_{it}^{w'}\beta - \eta_i^3}{\sigma_{\varepsilon_3}} \right) G\left(\frac{X_{it}^{e'}\delta + \eta_i^2 + \frac{\rho_{23}}{\sigma_{\varepsilon_3}}(w_{it} - X_{it}^{w'}\beta - \eta_i^3)}{\sqrt{1-\rho_{23}^2}}; \right. \\
&\qquad \left. \frac{X_{it}^{o'}\delta + \eta_i^1 + \frac{\rho_{13}}{\sigma_{\varepsilon_3}}(w_{it} - X_{it}^{w'}\beta - \eta_i^3)}{\sqrt{1-\rho_{13}^2}}; \; \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{13}^2)}} \right) \; ; \\
P(e_{it}^* \geqslant 0, w_{it}, r_{it}^{o*} < 0) &= \frac{1}{\sigma_{\varepsilon_3}} \phi\left(\frac{w_{it} - X_{it}^{w'}\beta - \eta_i^3}{\sigma_{\varepsilon_3}} \right) G\left(\frac{X_{it}^{e'}\delta + \eta_i^2 + \frac{\rho_{23}}{\sigma_{\varepsilon_3}}(w_{it} - X_{it}^{w'}\beta - \eta_i^3)}{\sqrt{1-\rho_{23}^2}}; \right. \\
&\qquad \left. \frac{-X_{it}^{o'}\delta - \eta_i^1 - \frac{\rho_{13}}{\sigma_{\varepsilon_3}}(w_{it} - X_{it}^{w'}\beta - \eta_i^3)}{\sqrt{1-\rho_{13}^2}}; \; -\frac{-\rho_{12} + \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{13}^2)}} \right) \; .
\end{aligned}
$$

$\Phi(x)$ and $\phi(x)$ are the standard normal cumulative distribution and density functions, respectively. $G(x_1, x_2, \rho)$ is the joint cumulative distribution function of the standard bivariate normal distribution with correlation ρ.

The unconditional likelihood function is obtained by integrating out the individual effects η_i:

$$
f^c(Y_i | X_i, \theta, \Sigma_\eta) = \int_{\mathbb{R}^3} f^c(Y_i | X_i, \theta, \eta_i) h(\eta_i; E_\eta, \Sigma_\eta) d\eta_i \; , \qquad \text{(C.3)}
$$

where $h(\eta_i; E_\eta, \Sigma_\eta)$ denotes the trivariate normal density function of $\eta_i = (\eta_i^1, \eta_i^2, \eta_i^3)$, with mean vector E_η and covariance matrix Σ_η, defined above.

We approximate this integral numerically according to the following procedure. First, we randomly simulate i.i.d draws $\eta_i^j = (\eta_i^{1j}, \eta_i^{2j}, \eta_i^{3j})$, $\forall j = \overline{1 \dots J}$, from the trivariate normal distribution H at a given value Σ_η (using a sequence of $J = 100$ Halton draws). Then, for each draw j we evaluate the conditional likelihood function $f^c(Y_i|X_i, \theta, \eta_i^j)$. Finally, we approximate the unconditional likelihood function for each individual as the mean of simulated f^c:

$$f^c(Y_i|X_i, \theta, \Sigma_\eta) = \frac{1}{J} \sum_{j=1}^{J} f^c \left(Y_i|X_i, \theta, \eta_i^j \right) . \tag{C.4}$$

The overall Simulated Maximum Likelihood function is then written as follows:

$$L = \prod_{i=1}^{N} \left[\frac{1}{J} \sum_{j=1}^{J} f^c \left(Y_i|X_i, \theta, \eta_i^j \right) \right] , \tag{C.5}$$

with the log-likelihood function:

$$lnL = \sum_{i=1}^{N} ln \left[\frac{1}{J} \sum_{j=1}^{J} f^c \left(Y_i|X_i, \theta, \eta_i^j \right) \right] . \tag{C.6}$$