The Effect of Age at Immigration on Employment Income

Ivelina Delahousse
2576279

Major Paper presented to the
Department of Economics of the University of Ottawa
Faculty of Graduate and Post-Doctoral Studies
In partial fulfillment of the requirements of the M.A. Degree

Supervisor: Miles Corak

ECO 7997
Ottawa, Ontario
April 2009

Table of Contents

List of Tables

List of Figures

Acknowledgements

This paper would not have been possible without the support of Statistics Canada in providing access to the Census data files and invaluable resources in the way of advice and documentation. Furthermore, many thanks must be extended to my motivating advisor, Miles Corak, who not only kept me on track but was an immense source of invaluable advice, inspiration, and encouragement.

And, last but not least, without the continued support from my family, especially my husband, Marc, the work for this paper would certainly have taken a longer time to complete.

1 Introduction

In countries like Canada, where immigrants make up large proportions of the population, there has been much interest in studying the determinants of migration and its effects on the economic and social development of these host countries. The well-being of immigrants and how well they integrate into their new environment has been of great concern for researchers and has been extensively studied on many levels, especially outcomes such as earnings[1]. The period of adaptation during which immigrants find their place in society and become settled depends tremendously on their language proficiency, experience, and background. This is well documented by Aydemir and Skuterud (2005), who quantify that the compositional shifts in language ability and region of birth are responsible for around one third of the observed deterioration in earnings. And, while they do not find declines in the returns to foreign education, they do find significant declines in the returns to foreign labour market experience, especially for men.

Being such an important factor for the integration of immigrants in their new environment, language ability and its effects on the economic outcomes have become the specific focus of many researchers (Chiswick and Miller, 2001, 2007, Bleakley and Chin, 2004, Casey and Dustmann, 2005). Immigrant children have gained particular attention in these studies because the prospect of a brighter future for them is a common motivation of many parents who decide to immigrate. Casey and Dustmann (2005), for example, find a significant association between the language proficiency of parents and that of their

[1] Baker and Benjamin (1994) studied the changes and differences of economic outcomes for Canadian immigrants from different cohorts. They find that entry earnings are lower for more recent immigrants. Bloom, Grenier, and Gunderson (1995) compare the earnings of Canadian immigrants to those of the Canadian-born population, and focus on the speed at which immigrant earnings grow over time compared to the Canadian-born. Aydemir and Skuterud (2005) investigate the causes of these deteriorations in entry earnings for male and female Canadian immigrants from successive cohorts.

children. Further, they find that the language deficiencies these children have are associated – at least for females – with poorer labour market outcomes.

There are not many studies of this sort. While the psychology literature may well cover the well-being of children as it relates to their development, the link between that literature and the well-being of children after migration is not as well covered. Fewer studies have concentrated on immigrant children and particularly how they grow and integrate into the society to which their parents brought them. Perhaps, this is in part owing to the data problems associated with this topic. Survey data on immigrants certainly exists and includes a wide range of economic indicators, the Longitudinal Survey of Immigrants to Canada (LSIC) offering the most recent and clearest example. Data on children is also available with significant amounts of information on child development, such as the National Longitudinal Survey of Children and Youth (NLSCY). But, data on immigrant children seems to be somewhat scarce, as none of these types of sources would be able to support a detailed study using the small sample sizes of immigrant children within them. Nevertheless, this is an important issue, especially for countries like Canada where the child population comes from a good mix of immigrant and non-immigrant families: according to the 2006 Canadian Census, 11.6% of children (under the age of 18) are born outside of Canada, and 19.6% are born to at least one immigrant parent.

If immigrant children are to be successful in becoming well integrated adults both socially and economically, then the level of social inclusion within their immediate surroundings will have a large impact on their degree of integration. In his study of the education and earnings mobility of second-generation Canadians, Corak (2008) finds that

social inclusion can have longer-term effects for immigrants, having importance even for their children's socio-economic progress. From this perspective, an evaluation of the adult outcomes of immigrant children would offer an assessment of how inclusive is a society.

This paper tries to fill some of the gaps in the literature with an analysis of the effect of age at immigration on the economic outcomes of adult immigrants. In particular, the paper focuses on employment income for adult immigrants (between the ages of 35 and 55) who arrived as children or young adults (at or before the age of 20), and explores discontinuities in the employment income function at points that coincide with the Critical Period Hypothesis, the idea that language acquisition occurs more easily and more fully if a child begins to learn a new language before a certain "critical" age. To deal with the data challenge, the 2006 Canadian Census of population is used, which offers the benefit of a large sample for the analysis.

The paper makes several contributions. First, it uses a model with elements of a regression discontinuity design to carefully explore discontinuities in the employment income function at points that may be connected to the Critical Period Hypothesis. The analysis is done for males and females separately and for several sub-populations classified by a language category variable, which is defined using the linguistic score of the immigrants' mother tongue. These language categories are meant to group together people coming from countries where the spoken language is of similar linguistic distance from English. The assumption is that it is more difficult to learn English for someone who speaks a language which is more "distant" from English than it is for someone who speaks a language which is "closer" to English. The ability to work with small sub-

populations is attributed to the large sample size of the data source. Second, the data itself is a contributing factor, as most available studies seem to be fairly dated. This paper, therefore, gives an up to date picture of the 'current' state of immigrant outcomes. Finally, it explores some of the consequences that can be associated with the Critical Period Hypothesis for the outcome of employment income.

The first main finding from this analysis is that discontinuities in the employment income function appear to be at different points for males and females. A negative effect for males appears to be at age at immigration of 12, putting employment income over $2,100 lower for males arriving at ages just after 12 compared to those arriving at ages just before and leading up to 12. For females, this effect appears at the age of 15 and is close to $1,400, representing a higher proportion of their average employment income than males. When the results are separated out by the language categories, differences in these effects are seen between the two genders within a category, as well as between the categories themselves. For males from countries where the spoken language is distant from English, the negative effect is as large as $5,600, representing over 13% of their average employment income. Furthermore, even though no significant effect is found for the group from countries where the spoken language is close to English, a surprising discontinuity of around $3,200 is captured for males from English speaking countries. The analysis of the further broken down language categories reveals that the effect within a language category can change sign with different origin countries, and that in many cases it is driven by the discontinuity of one country in that language group.

The rest of the paper is organized as follows. Section 3 presents a literature survey of some of the related work, offering an overview of the Critical Period Hypothesis and

how Chiswick translates it to an economic analysis. A detailed review of two main studies in this area using two very different approaches in their analysis of immigrant outcomes is also offered. The section ends with a short critique. Section 4 explains how the 2006 Canadian Census of population is used, and highlights some of the measurement issues related to the data. Some descriptive statistics and definitions of outcomes are also discussed in this section. The analytic methods are described in Section 5, followed by the results and analysis in Section 6. Finally, the paper concludes with Section 7.

2 Identification

With the increased availability of data and computing power, economists, especially those working on labour related topics, have been more and more interested in exploring and studying causal relationships stemming from policies or interventions, or from some individual behaviour. Consequently, empirical strategies and methods appropriate for such analysis have been the focus of much research. Angrist and Krueger (1999) extensively cover several commonly used approaches for causal inference models.

The main challenge with studies of causal relationships is to devise an identification strategy, the uncovering of a causal relationship from observational data. Quasi-experimental methods offer a motivating example for this discussion. These studies investigate causal relationships through the comparison of two counterfactual states (Angrist and Krueger, 1999). For example, in the case of studying the effects of a new policy on a certain outcome, one possible approach would be to compare the outcome for a group of people affected by the policy (the treatment group) to the outcome for a group of people with similar characteristics not affected by the policy (the control group). Then, the differences observed between the two outcomes could be interpreted as the causal effect, as long as all the necessary variables are included in the model.

There are three key factors that validate this approach. The first is that the two groups being compared are otherwise identical with the only difference being the exposure to the policy for one group but not for the other. The second point is that there are no variables omitted from the model which are correlated in some way to the other variables of interest (Angrist and Krueger, 1999). The third condition refers to common

trends in the underlying evolution of the variable of interest. When these conditions are met, and an appropriate econometric technique is used, then the causal relationship of interest can be "identified" (Angrist and Krueger, 1999). If one of these elements is lacking, then the causal relationship cannot be properly explained by the variables included in the model.

Outlined in this section are two empirical strategies which try to overcome these types of problems. Instrumental Variables (IV) methods are often the first option researchers consider as an appropriate method for a causal inference study. Moreover, the Regression Discontinuity (RD) design has become a more commonly used technique, particularly in studies focusing on intervention or policy effects.

2.1 Instrumental Variables (IV)

Instrumental Variable (IV) methods are commonly relied on when dealing with endogeneity in a model. It is also used to correct for omitted variables bias and measurement error. Endogeneity refers to the presence of an endogenous variable (that is, a variable whose value is determined through its relationship with other variables) in the set of explanatory variables. Endogenous variables are as a result correlated with the error term. For example, the causal relationship between language proficiency and earnings cannot be estimated through a simple Ordinary Least Squares (OLS) regression because some immigrants with higher proficiency might have some special characteristic or be affected by some other factor which has positive effect on their proficiency levels. As a result, OLS estimates of the effect of language proficiency on earnings would be biased since language skills are an endogenous variable. Language proficiency is correlated with the error term by virtue of the fact that while it may be correlated with earnings, there is

also some other factor that is omitted from the regression which also explains earnings in some way through its effect on language proficiency.

Provided that a suitable instrument exists for the endogenous variable in question, an IV strategy would remove the resulting bias found in the corresponding OLS estimates. To implement this approach, however, an "instrumental" variable must be available, such that, it is correlated with the endogenous variable, and correlated with the outcome variable of interest only through its relationship with the endogenous variable (that is, it is uncorrelated with the error term in the regression of interest). If such a variable does exist, then the endogenous variable is first regressed on the instrument, then the fitted values of this regression are entered into the equation of interest in the initial place of the endogenous variable. Because this is a two-step procedure of OLS regressions, it is often referred to as Two-Stage Least Squares (2SLS).

For example, if the equation of interest is $y_i = \alpha + \beta x_i + \gamma W_i + \varepsilon_i$, where y_i is the outcome variable of interest, and x_i is the endogenous variable, with W_i being a set of other exogenous explanatory variables, then for a set of instruments for x_i, say Z_i, the 2SLS model would be set up in the following way. In the first stage, the endogenous x_i would be regressed on its instruments: $x_i = \alpha_1 + \beta_1 Z_i + \gamma_1 W_i + \varepsilon_{1i}$. Then, in the second stage, the fitted values from the first stage regression would be included in the regression of interest replacing the endogenous variable: $y_i = \alpha + \beta \hat{x}_i + \gamma W_i + \varepsilon_i$. The relationship between y_i and x_i is now explained through the part of the variability in x_i that is uncorrelated with the omitted variables in the regression, which removes the initial bias resulting from the fact that the effect from x_i reflected the total variability in x_i (Angrist

and Krueger, 2001). Bleakley and Chin (2004) implement this strategy for the causal relationship of language skills on earnings using age at immigration interacted with a dummy for non-English-speaking country of origin as the instrument.

This approach, however, strongly depends on the validity of the instruments being used. Angrist and Krueger (2001) warn that a weak instrument (that is, one "that is correlated with the omitted variables" or one that is only weakly correlated with the endogenous variable) can result in estimates having bias that is greater than the corresponding bias associated to the OLS estimates. Furthermore, they point out that while IV estimates "are consistent, they are not unbiased" (that is, the sampling distribution of the IV estimator is not centered on the parameter of interest for any sized sample), so studies using these methodologies should rely on the use of large samples (Angrist and Krueger, 2001).

2.2 Regression Discontinuity (RD)

Regression discontinuity (RD) analysis is generally used for studying the causal effects of interventions (or binary treatments). It is based on the idea that the value of a covariate determines the assignment of the treatment by having a value on either side of a fixed threshold. Then, as long any association between this covariate and the outcome variable is smooth, if a discontinuity is observed in the conditional distribution of the outcome as a function of this covariate at the cut-off point, then it is interpreted as a causal effect of the treatment (Imbens and Lemieux, 2008). This design is regarded as having validity, but only internally, since the treatment effect is estimated locally to the critical point (Nichols, 2007).

There are several assumptions which define a RD design which are well summarized by Nichols (2007). The first is that there exists a discontinuity at some fixed threshold (the critical point) of the "forcing" variable which defines the level treatment. The second is that the treatment is not randomly assigned, but is determined by an observable variable (the forcing variable), having a value on either side of some fixed threshold. The third assumption deals with the characteristics of the observations on either side of the threshold. These two sets of observations must be otherwise identical with only the treatment variable being different. In other words, individuals cannot affect whether they fall on one side of the threshold or the other in order to place themselves as part of the treatment group or as part of the control group. Finally, the last essential assumption for a RD design is that conditional on the treatment, all other variables related to the outcome variable are smooth functions of the variable which determines the treatment. This requirement ensures that the outcome variable jumps at the cut-off point only because of the discontinuity in the treatment changing from zero to one (Nichols, 2007).

In their practical guide to using and implementing RD models, Imbens and Lemieux (2008) argue that graphical analyses should be an integral part of all RD models because graphs can nicely illustrate the discontinuity in question and provide a visual representation of the identification strategy. This is, in a sense, the starting point for determining how appropriate is the RD design as an identification strategy for a particular study and in testing if the above-mentioned assumptions are satisfied.

Imbens and Lemieux (2008) list three types of plots that should be examined for a RD design. The first is, of course, a plot of the mean of the outcome variable against

different values of the forcing variable. With this graph, the first two assumptions of the RD model can be verified. If no discontinuity is observed in the graph at the critical point, then it is unlikely that the model will detect a statistically significant jump at that point. On the other hand, if a discontinuity is observed, it would validate the first assumption. Furthermore, it is also important to confirm at this point that there do not appear to be discontinuities at other points in the graph. If there are indeed other jumps which cannot be explained, then the causal interpretation at the critical point may not be all that credible as it would suggest that the treatment could be assigned based on different values of the forcing variable. Thus, observing a single point of discontinuity would validate the second assumption that the treatment is determined by the forcing variable precisely at the critical point of interest.

The second assumption is verified using a plot of the density of the forcing variable, which is crucial in showing whether a discontinuity in the distribution of that variable exists at the critical point. If a discontinuity is in fact observed, then the design can be discredited as this would suggest that the value of the forcing variable may be manipulated by the individual resulting in a non-random assignment of the treatment. Having observations which are otherwise identical on either side of the critical point is fundamental to the RD design, and, to be able to construct a solid RD model, great care must be taken to show that there is no evidence of self-selection by individuals. Therefore, this part of the graphical analysis is of importance.

Finally, the second important plot is a plot capturing the relationship between other covariates and the forcing variable. This plot is useful in determining if there are variables other than the forcing variable which could possibly be responsible for the jump

at the critical point. Examining such a graph would address the fourth assumption in ensuring that all other covariates of the outcome are smooth functions of the forcing variable.

This design comes in two flavours: Sharp Regression Discontinuity (SRD) and Fuzzy Regression Discontinuity (FRD). The sharp design is used when the treatment variable is deterministically defined by the forcing variable (that is, the level of treatment jumps from zero to one precisely at the cut-off point), whereas the fuzzy design is used when the expected value of the treatment, conditional on the forcing variable, increases discontinuously at the cut-off point (Nichols, 2007).

2.3 RD design as an IV setup

The implementation of a fuzzy RD design leads to an IV model setup (Angrist and Pishke, 2009), which can be easily estimated with a 2SLS model. Jumps in the probability of treatment are used as the source of identifying information, with a variable indicating the point where the probability of treatment is discontinuous acting as the instrument. For example, if T_i is the treatment variable and $P(T_i = 1 \mid x_i)$ is the probability of treatment conditional on the forcing variable, then the point where this probability is discontinuous can be used as an instrument for T_i. Then, in the first stage, T_i is regressed on this instrument, together with a functional form of the forcing variable. And, in the second stage, the results from the first stage regression are included in the place of the treatment variable in the equation of interest.

Designs specified in this way have many applications for economists. For example, Lemieux and Milligan (2008) use fuzzy regression discontinuity to estimate the effects of social assistance on labour market outcomes resulting from a policy in the province of

Quebec. Angrist and Lavy (1999), on the other hand, use this method to study the causal impact of class size on test scores, where class size is determined by a rule, "Maimonides" rule.

It should also be noted that the sharp RD design differs from the IV strategy in that the assignment variable not only defines the treatment, but it can also have a direct impact on the outcome. In the IV scenario, the instrument affects the endogenous variable in such a way that exogenous variation can be isolated and used to estimate the causal effect on the outcome; thus, the instrument has only an indirect impact on the outcome. It is, therefore, the sharp RD design that motivates the analysis in this paper.

3 Literature Survey

3.1 The Critical Period Hypothesis

The Critical Period Hypothesis (CPH) comes from the psychological literature and is based on the idea that there is a "critical" period beyond which second language acquisition is much more difficult and native-like competencies are unlikely to be achieved. This hypothesis initially stemmed from biological research on brain maturation and later was linked to research on cognitive abilities. Penfield and Roberts (1959) were the first to introduce this idea to the literature. A crucial component of the CPH is that there exists a distinct discontinuity in the decline of the ability to learn a second language which would define the critical period. Researchers have had difficulty arriving at a consensus on what age defines the CPH. Some have suggested specific teen years, while others have considered the school age as the breaking point.[2] Furthermore, in testing for the existence of a critical period, many have found that proficiency declines more or less continuously with the age at arrival without there being any specific discontinuity to indicate the end of one period and the beginning of another (Hakuta *et al.*, 2003).

Despite the lack of agreement among researchers on the existence of a critical period or on how to define it if one does exist, studies on the CPH have shown that there is an important relationship between age and second language acquisition. And since language proficiency among immigrants is a key indicator of how successfully they integrate into their new society, the CPH has captured the attention of economists, such as Chiswick, who have given it economic meaning. The effects of the age at which

[2] Hakuta *et al.* (2003) review various literature on the CPH from both sides of the debate: that there is a distinct age which defines the critical period for learning a second language; or that second language learning becomes more difficult with age due to other external factors such as education, as well as cognitive aging.

someone begins learning a second language on how successful they are at mastering it has large implications for the well-being of immigrants and, more specifically, their children. This is precisely the focus of some of Chiswick's work.

3.2 Chiswick's work

Chiswick sees language proficiency as a form of human capital. It fits the definition of human capital since being proficient in the local language is something that is productive for the labour market and for the society as a whole; it is embodied in the individual; and, it comes at a cost to the individual (Chiswick, 2007). To understand language proficiency better, Chiswick studied the determinants and the consequences of being proficient using data from four countries.

In studying the determinants of language proficiency, he focused on three aspects of language acquisition: exposure to the new language, both prior to and post migration; efficiency in learning a new language; and the economic incentives that exist for learning the new language (Chiswick, 2007). He refers to these as "the three E's". In terms of exposure, he found that proficiency increased with years since migration but decreased if the immigrant lived in a highly concentrated linguistic enclave. Proficiency also increased for immigrants with pre-migration exposure to the language (e.g. former English colonies). With respect to efficiency, his results showed that proficiency increased with the schooling level, but decreased with greater ages at migration. In terms of family environment, marriage before or after migration also showed to be important for the proficiency of immigrants. Those with the highest proficiency were ones who were married after migration, those who were not married were less proficient, while those married before migration were the least proficient. While the presence of children

had no impact on the father's proficiency, it had a negative effect on that of the mother (Chiswick, 2007).

In studying the consequences of language proficiency, Chiswick (2007) found that greater proficiency translates into greater earnings and immigrants who live in concentrated linguistic/ethnic enclaves experience lower earnings than those who do not. These kinds of results brought his attention to the CPH and the debate in the literature around it, arguing whether or not there exists a certain "critical" age, after which it is too difficult for an immigrant to become fully proficient in a second language. Chiswick and Miller (2008) used a rigorous approach to identify if a critical period exists for the age at arrival. Their model uses the 2000 US Census data and is constructed as a probit regression with proficiency being the dependent variable explained by variables representing the concepts described by the three E's (exposure, efficiency, and economic incentives).

The findings show that proficiency (in English) monotonically declines with age at migration for all the groups that were considered and while some groups had steeper declines at first compared to more gradual ones, the patterns observed were similar across gender within the same country of origin. When compared to the reference group of those arriving at very young ages (0-1 years old), age at migration was found to have a statistically significant effect on proficiency at the 5% level for ages sufficiently far from the reference category (around the ages of 5 to 10). However, when consecutive age categories were compared, none showed to be significantly different. Furthermore, the same analysis was done separately for immigrants with mother tongues close to English and for those with mother tongues distant from English. Again, the patterns are consistent

with those previously described, except that for the group with mother tongues distant from English, the decline in proficiency happens faster with higher ages at migration compared to the group with mother tongues close to English. Chiswick and Miller (2008), therefore, concluded that even though their analysis was not able to identify a "critical" age, it was able to show the fundamental relationship between proficiency and age at migration and that there can be significant differences for ages of migration sufficiently far apart. Their results also emphasized that the concept of linguistic distance plays an important role in the analysis of language proficiency.

Linguistic distance refers to the measure of the difficulty of learning a language. Although it may seem that measuring how difficult it is to learn a language cannot be done in a quantitative way, Chiswick and Miller (2005) developed such a measure for the distance between English and other languages. The general idea behind this measure is that if English speaking people have more difficulty learning one language over another, then the more difficult language is more distant from English than is the easier one. A language was considered to be more difficult to learn than another language if after learning it for a set period of time, the English speaker was less proficient in that language than in the other language (having spent the same amount of time learning it). Results from standardized tests on proficiency and indices developed by linguists were used to create the linguistic distance measure (Chiswick and Miller, 2005).

Applying this measure in the analysis of proficiency using 1990 US and 1991 Canadian Census data, Chiswick and Miller (2005) found that the greater the distance between the mother tongues of the immigrants and English, the lower the level of their English proficiency, other variables constant. The results from the Canadian data show

how strongly linguistic distance affects proficiency. After 15 years of living in Canada, of the immigrants with the most distant mother tongues, 10% could not carry on a conversation in English or French and only 5% of them usually spoke English or French at home. This is in contrast with the experience of immigrants with the least distant mother tongues to English, of whom after 15 years, only 1% could not carry on a conversation and 58% usually spoke English or French at home (Chiswick and Miller, 2005).

Using a more sophisticated model where language proficiency, seen as a type of human capital, is explained by more variables than just distance, Chiswick and Miller (2001) used the Canadian Census data of 1991 to study the determinants of language proficiency among male immigrants. Their analysis is designed around a multinomial logit model with a trichotomous dependent variable of language proficiency explained by a set of variables representing the effects of the three E's (exposure, efficiency, and economic incentives) and of wealth.

They make three main contributions to the literature at the time of their study. First, to represent the country of birth in the model, they used five variables to reflect several dimensions of the origin country. These variables (geographic distance, linguistic distance, minority language concentration, refugee status, and former colony) are used in part due to limitations in the data available on country of birth, but more importantly to reflect the effect of the three E's and give a more behavioural interpretation to the factors affecting language proficiency. The inclusion of the linguistic distance and the minority language concentration in the model are their other two major contributions (Chiswick and Miller, 2001).

As expected, the results of the study show that age at migration, educational attainment and duration of residence in Canada are all significant determinants of language fluency. Of the five variables representing the effects of birthplace, linguistic distance, minority language concentration, and refugee status have a negative effect on language proficiency. This means that having an origin language which is very far away from English, living in a highly concentrated ethnic community, or being a refugee contributes to a lower probability of being fluent in the new language. On the other hand, coming from a country which is of greater distance from the host country and coming from a former colony have a positive effect resulting in greater probabilities of having higher proficiency (Chiswick and Miller, 2001).

3.3 Earnings Literature

Bleakley and Chin (2004) studied the causal effect of language skills on earnings using an instrumental variable (IV) strategy with the interaction between age at arrival and a dummy for non-English-speaking countries of origin acting as the instrument. With this approach, the non-language effects of age at arrival (which could also affect earnings) can be separated from the language effects. The intuition is that since immigrants from non-English-speaking countries are faced with all the same things as immigrants from English-speaking countries with the one addition of the new language, then any differences in earnings between earlier and later arrivers from non-English-speaking countries in excess of the differences observed for the immigrants from English-speaking countries can be interpreted as the language effects.

The data used by Bleakley and Chin (2004) is the 1990 US Census Public Use file. They restrict their sample to people who have lived in the US for 16 to 30 years and have

arrived before the age of 18. The age at arrival variable is defined using different time intervals and the language proficiency variable is self-reported to one of four categories. These broad definitions are certainly a limitation imposed by the data which affect the estimates of the model parameters.

Motivated by theory and research on language acquisition, Bleakley and Chin (2004) first treat age at arrival as a binary instrument (interacted with the dummy for non-English-speaking country of origin). Following psychobiological literature and the CPH, they define their binary dummy for age at arrival using the age of 12 as the critical point. This model explicitly incorporates the CPH, and therefore, is founded on the notion that there is something particular about the teenage years that changes the way in which a child learns a language and that this (together with country of origin) is what explains the level of language proficiency of immigrant children as adults. The results from this formulation of the model show that the non-language effects are very small and suggest that language skills account for most of the effects on wages.

In an extended version of the model, Bleakley and Chin (2004) redefine age at arrival to permit variation in language-learning ability. They include a set of country of birth fixed effects, γ_j, a set of age at arrival fixed effects, δ_a, and a set of (exogenous) explanatory variables, w_{ija}, where i identifies the individual, j refers to the country of birth, and a represents the age at arrival. This specification demonstrates the progressively poorer language abilities of immigrant children from non-English-speaking countries the older they are when they first immigrate. In the first stage, language skills (the endogenous variable), x_{ija}, are related to the instrument, k_{ija}, which is defined as $k_{ija} = \max(0, a_i - 11) \times N_j$, where a_i is individual i's age at arrival and N_j is the non-

English-speaking country dummy variable. In the second stage, wages (the outcome), y_{ija}, are related to language skills.

Equation 1: Bleakley and Chin 1st stage equation

$$x_{ija} = \alpha_1 + \beta_1 k_{ija} + \delta_{1a} + \gamma_{1j} + w'_{ija}\rho_1 + \varepsilon_{1ija}$$

Equation 2: Bleakley and Chin 2nd stage equation

$$y_{ija} = \alpha + \beta x_{ija} + \delta_a + \gamma_j + w'_{ija}\rho + \varepsilon_{ija}$$

The results show that with respect to earnings, language skills seem to be paying off well: those who speak English well earn 33% more than those who speak poorly; and those who speak very well earn 67% more.

However, the authors question the validity of their chosen "control" group of immigrants from English-speaking countries as they recognize that it is possible the effects of age at arrival experienced by immigrants from non-English-speaking countries may be the result of some factor other than language which is related to age at arrival but has been omitted. Furthermore, they find that when comparing their IV results to what results OLS would have produced, the OLS estimate seems to be downward biased, which is in contrast to what would be expected. If it is ability that is responsible for the endogeneity of language skills, then this would imply that the OLS estimates should be larger than the IV estimates. Consequently, to validate their approach, Bleakley and Chin conduct a series of robustness checks.

First, they consider two alternative explanations for the differences between the age at arrival effects of the treatment and control groups which are unrelated to the effects of language. One hypothesis is that because richer countries tend to have better schooling and non-English-speaking countries tend to be poorer, then the observed effects would also be reflecting the returns to schooling in the origin country of the immigrants. In

controlling for this possible effect by including some origin-country school quality indicators in the model, the final results do not come out to be all that different. The second alternative hypothesis is that unlike immigrants from English-speaking countries, immigrants with children from non-English-speaking countries may be considering their children's age in their immigration decision. In other words, knowing that their children would have a language barrier to overcome, they may be purposely immigrating when their children are younger. If this is the case, then the observed effects would also reflect these decision characteristics. By looking at the age at arrival distributions for the group of immigrants from non-English-speaking countries and for the group from English-speaking ones, Bleakley and Chin do not find that decisions of this kind are taking place among the non-English-speaking immigrants.

Finally, the authors (Bleakley and Chin, 2004) study the possible reasons for having obtained a smaller OLS estimate compared to the IV estimate. They focus on how the measurement error in the language skills variable affects the OLS and IV estimates. With the help of data from the National Adult Literacy Survey (NALS) containing literacy test scores which can indicate "true" language ability, Bleakley and Chin find that when they allow for classical, or even non-classical, measurement error, their results still show that the effect of language skills on earnings is significant. When they remove the biases associated to the measurement error, the OLS estimate becomes higher than the IV estimate by 10 to 20 percent. Given that the IV estimate accounts for the endogeneity of language skills whereas OLS does not, they conclude that the remaining differences between the estimates can be attributed to that fact (Bleakley and Chin, 2004).

Using a very different approach, Schaafsma and Sweetman (2001) also study the effect of age at immigration on earnings, but they focus on explaining why age at immigration matters for earnings. Rather than comparing one group of immigrants to another group of immigrants (as do Bleakley and Chin, 2004), using Canadian data[3], they compare immigrants to their Canadian-born counterparts from a similar profile and observable characteristics. In their two step approach, they first estimate the age-earnings profile for the Canadian-born sample; then, they estimate a model to determine the effect of age at immigration on the difference between observed immigrant earnings and those predicted from the Canadian-born age-earnings profile. The fundamental assumption they make is that the effect of age on earnings is the same for immigrants as for the Canadian-born. This is how they identify their model.

To estimate this model, Schaafsma and Sweetman (2001) use the Canadian Census Public-use microdata files of 1996, 1991, and 1986 excluding some geographic areas where the characteristics of immigrants are severely grouped. As is done by many others, they also restrict their analysis to males between the ages of 16 and 64 who were employed. With these data and the two-stage estimation strategy described above three sets of regression results were produced.

The first set of regressions (one for each Census year) focus strictly on the effect of age at immigration and year of immigration on the dependent variable as defined by the estimation strategy[4]. The 1996 results from these regressions show that an average immigrant who arrived between the ages of 45 and 64 received 32% less than one who arrived between the ages of 0 and 4. The monotonic decline of earnings with age at

[3] This is the only study that was found to use Canadian data for this type of research.
[4] The dependent variable is the difference between the observed immigrant earnings and the predicted Canadian-born earnings from the age-earnings profile.

immigration observed in other studies is not seen here for any of the three datasets. On the contrary, immigrants who arrived in their late teens appear to have lower earnings than both those who arrived at a younger age and at an older age.

In the second set of regressions controls are added for observable differences between immigrants and the Canadian-born. The earnings gap observed here is similar to the one from the first set of regressions but the effect seen for those who immigrated in their late teens is no longer significant. Finally, in the third set of regressions, the returns to observable characteristics are allowed to vary by adding control variables to both stages of the model framework.

There are two other studies relevant to this discussion, which focus on immigrant earnings from the perspective of the returns to education and literacy skills. Ferrer and Riddell (2008) use the Canadian Census Public-use files (1981 – 2001) to investigate sheepskin effects[5] for immigrants. They find that while the returns to years of schooling and experience are lower for immigrants compared to their Canadian counterparts, when it comes to having a degree, the story is different. Immigrants with a degree benefit from higher earnings compared to immigrants without one, the earnings gains being at least as large as those of Canadians with a similar degree. This analysis is also done for immigrants who arrived in Canada before the age of 20. Ferrer and Riddell find that differences in the returns to education between this group and the Canadian-born are not nearly as pronounced as those documented for immigrants arriving as adults.

Ferrer, Green and Riddell (2006) use survey data to study the effect of literacy on immigrant earnings. Their findings show that literacy skills among immigrants are

[5] Sheepskin effects refer to the increase in earnings which is attributed to the completion of a degree when years of schooling are taken into account.

substantially lower than those among the Canadian-born. But, the findings did not show immigrants to have lower returns to these literacy skills compared to similar Canadian-born individuals. However, the authors did find that literacy skills have an important effect on earnings, suggesting that the lower levels of literacy among immigrants can explain at least part of the observed differences in earnings with the Canadian-born (Ferrer, Green, and Riddell, 2006).

3.4 Critique

In an effort to further support their results, Bleakley and Chin (2004) conduct a large set of evaluation studies for robustness, contribution of education to the effect of language skills on wages, the degree of bias on the regression coefficients, and measurement error associated to the language variable. However, despite all the care that is taken to justify and support these methodologies, it is still not clear that age at migration and country of origin are the only variables that affect language proficiency. The work done by Chiswick and Miller (2005) on linguistic distance provides a different view of the analysis on language proficiency and demonstrates that language distance has a clearly strong effect on language proficiency. In their results, immigrants from countries of origin with mother tongues that are very different (distant) from English show to be much less proficient than immigrants from countries with mother tongues that are closer to English. Including a measure for language difficulty of this type may, therefore, be important.

In Chiswick and Miller's study (2001) of the determinants of language proficiency, the analysis is focused on males only. This is typical in many studies because males are considered to be stable participants in the labour market. However, given the importance

of the results obtained for males, and the evidence from the work of Casey and Dustman on the dramatic difference in the effects on economic outcomes between males and females, it seems that separate analysis by gender may be of value.

This paper, therefore, tries to address some of these gaps. First, it uses an innovative method to model and explore discontinuities in employment income. Motivated by Chiswick and Miller's (2008) work in testing the CPH, this study investigates the existence of discontinuities locally around the point in the employment income function which would coincide with the CPH. Relying on elements from a regression discontinuity framework, the model is designed to not only test if breaks in the employment income function do exist at points related to the CPH, but also to capture and quantitatively measure these discontinuities.

To do this, a regression model is developed based on the specification of a sharp RD design. The outcome of interest (employment income) is regressed on a treatment variable, defined as a dummy representing the critical point of interest, and the variable for age at immigration. The estimates of the treatment effect are used to identify and measure existing discontinuities at points that coincide with the CPH. It should be noted, however, that the RD design is used only as a tool to define and structure the model in this study. It is not used as an identification strategy to indicate causality. Furthermore, this study does not directly look at language proficiency and it relationship to employment income, but is rather motivated by the CPH to explore discontinuities in the employment income function at points that may be related to the CPH.

The analysis is done separately by gender so that the effects would reflect closely the impact for males and females. Results are also produced for different language

categories representing measures of linguistic distance. This allows for varying effects between individuals from countries where the spoken language is close to English and others where the spoken language is more distant from English.

4 Data and Measurement Issues

4.1 Census 2006 data: Advantages and Limitations

The Canadian Census of population is conducted every five years and collects a large variety of information on the Canadian population. In 2006, a random sample of 20% of households received the long form questionnaire which covers many detailed variables about all aspects of a person's social and economic status.

This data source has several advantages over survey data. First and foremost, its large sample size permits analyses to be done at a detailed level, or with a focus on very precise and small domains. Second, unlike many surveys that target very specific populations, the Census targets the entire population, so every population group is well represented.

These are important factors for this study as the focus here is on a particular set of immigrants (those between the ages of 35 and 55 who immigrated as children or young adults) and the analysis is concentrated on a certain age at immigration. Thus, the large sample size is not only crucial here, but it also comes with the advantage of being able to further examine even smaller domains such as specific immigrant communities.[6]

In spite of its advantages, the Census also has certain limitations and notes of caution. Regardless of how much effort is put toward obtaining data of the highest quality, it is inevitable that the estimates in the end will be subject to a certain degree of error. This error can come from various sources and can be more or less important depending on the source and the prevalence of it. There are two types of error which are

[6] The immigrant communities are defined using the language score associated with the language spoken in the country of birth for the immigrant. These scores are taken from Chiswick and Miller (2005) and mapped to the languages found in the dataset used here. For details on the mapping, refer to Section 4.5.

of concern for this study: response errors (or sometimes known as measurement errors), and processing errors. Response errors occur when the respondent gives an incorrect response due to a misinterpretation of the question or incorrectly records the answer, sometimes without even being aware of it. Processing errors occur during data capture, response coding, or the production of derived variables (Statistics Canada, 2007).

The 2006 Canadian census contains some known errors related to several of the immigration variables. It is important to discuss this here, as these errors will directly affect the analysis that follows. The derivation of one of the most critical variables for this study, the age at immigration variable, was significantly affected by a processing error. During the creation of this variable, an estimation process was somehow omitted which resulted in an increase in the estimate of the age at immigration by one year for some records. Consequently, the estimate of the median age at immigration for the total immigrant population is one year higher[7] than what the correct estimate should be (Statistics Canada, Place of Birth, Generation Status, Citizenship and Immigration Reference Guide, 2008).

4.2 Sample Selection Rules

The focus of this study is on measuring discontinuities in the adult outcomes for immigrants who arrived at a young age, which capture the effect of age at immigration on the employment income function. The selected sample is, therefore, restricted to the

[7] The effect of this error can be seen in the distribution of the age at immigration for a group of individuals of a particular age. A certain proportion of these individuals have an age at immigration which is higher than their actual age. Since the sample of immigrants here is restricted to those who are 35 to 55 years old and who have come to Canada at or before the age of 20, it was not possible to get a sense of the magnitude of this error for this set of individuals because there is nothing to indicate that while their age at immigration was coded to 12, it should really have been coded to 11. However, for the full set of 35 to 55 year old immigrants (unrestricted to a time of arrival), overall 0.5% were found to have an age at immigration higher than the actual age.

group of adults who are between the age of 35 and 55 in the 2006 Canadian Census of population Master file, representing approximately 10 million Canadians. For this group, the outcome of interest, employment income, can provide a fairly good measure of permanent income because these individuals are generally at a similar point in their life cycle earnings profile. Among the immigrants in this group, only those who were 20 years of age or younger when they immigrated and those who lived in one of the 10 provinces were considered as part of the sample. Furthermore, individuals with extremely high employment income[8] or with negative employment income were also excluded from the sample. This exclusion was done to ensure that a few outlying observations did not influence the mean employment income for a particular group to be higher or lower potentially resulting in a misperception of the observed effects.

Initially, it was considered to keep the refugee population in the sample, as they would have provided good exogenous variation since their choice of when to immigrate is not based on the age of their children. However, those individuals did not have a value for the age at immigration variable on the data file because they do not have landed immigrant status. The age at immigration variable is only available for immigrants with landed immigrant status or citizenship because it is derived using the year of becoming a landed immigrant as the reference point. Refugees, of course, are not landed immigrants and thus had to be excluded from this analysis.[9] What this means, though, is that it is not possible to know whether a given individual immigrant initially came as a refugee and then subsequently received landed immigrant status and potentially even citizenship. So, for all immigrants who came as refugees, the age at immigration variable in a sense

[8] Employment income was considered to be "extremely high" when it was more than 3 standard deviations larger than the overall mean employment income.
[9] The age at immigration variable is given a code of -4 for Refugee individuals in the data.

overstates their true age at immigration since it would not capture any time spent in the country prior to becoming a landed immigrant. This is another source of measurement error. Institutional residents were also removed from the sample because no income information is available for this part of the population.

In the process of selecting the sample, some cases with rare characteristics were discovered. Some individuals who were coded as Canadian-born in one variable (the age at immigration variable[10]) were also coded as being first generation immigrants in another variable (the generational status variable). Similarly, some others were coded as both Canadian-born and as having an age at immigration. This can occur in some rare situations. For example, a mother giving birth to her baby in Canada can move back to her country of origin shortly after. Then, if the child immigrates at a later date, he or she would legitimately have an age at immigration despite having been born in Canada. There were just over 5,900 such cases representing less than 29,000 people. These cases were excluded from the sample as it is unclear whether they should be considered as part of the immigrant group or as part of the Canadian-born group. Table 1 gives the sample counts and the corresponding population size of the selected sample and all excluded portions.

Overall, these sample selection rules yielded a sample size of just over 136,000 adult immigrants representing almost 695,000 people who immigrated at the age of 20 or younger and live in the ten Canadian provinces.

[10] The age at immigration variable was given a code of -3 for the Canadian-born units in the data.

Table 1: Sample Selection

Description	Sample Count	Population Total
Total (adult population of age 35 to 55)	**2,137,809**	**10,225,084**
Total population – EXCLUDED	**74,629**	**151,411**
NA – Institutional residents	53,111	53,111
Refugees	15,610	69,766
Cases with rare characteristics	5,908	28,534
Total population (after exclusions)	**2,063,180**	**10,073,673**
Total living in the territories – DROPPED	20,333	31,408
Generation 3 – Living in the ten provinces – DROPPED	1,315,123	6,321,972
Generation 2 – Living in the ten provinces – DROPPED	258,099	1,302,630
Generation 1 – Living in the ten provinces (arrived after the age of 20) – DROPPED	332,559	1,717,506
Generation 1 – Living in the ten provinces (arrived at or before the age of 20 with employment income of outlying or of negative value) – DROPPED [see note b.]	1,053	5,347
Generation 1 – Living in the ten provinces (arrived at or before the age of 20 with employment income not of outlying or of negative value) – SELECTED	**136,013**	**694,810**

Notes:

a. First generation (immigrant) status, Generation 1, comes from a derived variable representing the generational status of an individual and means that the individual is born outside Canada; thus, "Generation 1" refers to immigrants. Similarly, "Generation 2" refers to individuals born in Canada to at least one foreign-born parent, and "Generation 3" refers to those born in Canada to two Canadian-born parents.

b. Individuals with employment income of outlying value were removed from the sample so that they would not be responsible for pushing the average employment income upwards for specific ages at immigration resulting in a perceived discontinuity of large value when in reality the magnitude may be much smaller or even insignificant. An individual's employment income was considered as outlying if it was greater than three standard deviations away from the mean employment income for the overall population of immigrants. Individuals whose employment income was negative were also removed from the sample since this is probably a reflection of temporary losses for self-employed individuals and is not a good indication of their permanent income.

4.3 Descriptive Statistics

Some of the characteristics describing this population are discussed in this section. Throughout the following set of analyses, it is important to remember that the set of immigrants being studied are not recent immigrants, or even immigrants who necessarily chose to immigrate to Canada. They are immigrants who immigrated as children or as young adults, and therefore have for the most part followed their parents in their immigration decision.

The age distribution of this sample of immigrants is given in Table 2. This is an important characteristic to examine for the current analysis because employment income

can vary with a person's age. The distribution is quite clearly not uniform having what appears to be a small hump around the ages of the early 40's followed by a dip around the late 40's. This seems to be the case for the overall population, as well as for each gender. The distribution between the two genders also does not appear to be quite equal. More females seem to make up the population of ages from 35 up to the early 40's, after which, it is males that make up the larger portion.

Table 2: Age distribution

Age	Sample Count			Population Distribution		
	All	Females	Males	All	Females	Males
35	6,722	3,424	3,298	5.0%	4.9%	5.1%
36	6,513	3,260	3,253	4.8%	4.7%	5.0%
37	6,555	3,330	3,225	4.9%	4.8%	5.0%
38	6,494	3,354	3,140	4.8%	4.8%	4.9%
39	6,857	3,544	3,313	5.1%	5.1%	5.1%
40	6,882	3,563	3,319	5.1%	5.1%	5.1%
41	7,310	3,720	3,590	5.4%	5.3%	5.5%
42	7,342	3,735	3,607	5.4%	5.3%	5.4%
43	7,089	3,529	3,560	5.2%	5.0%	5.4%
44	6,832	3,477	3,355	5.1%	5.1%	5.1%
45	6,787	3,507	3,280	4.9%	4.9%	4.9%
46	6,180	3,200	2,980	4.5%	4.5%	4.5%
47	5,675	3,021	2,654	4.2%	4.3%	4.0%
48	5,646	2,966	2,680	4.1%	4.2%	4.1%
49	6,064	3,152	2,912	4.5%	4.5%	4.4%
50	6,241	3,336	2,905	4.5%	4.7%	4.3%
51	6,277	3,267	3,010	4.6%	4.7%	4.5%
52	5,995	3,110	2,885	4.4%	4.4%	4.3%
53	6,042	3,169	2,873	4.3%	4.3%	4.3%
54	6,050	3,176	2,874	4.4%	4.5%	4.3%
55	6,460	3,336	3,124	4.7%	4.7%	4.7%
Sample Count and Population Total	**136,013**	**70,176**	**65,837**	**694,810**	**358,190**	**336,619**

When higher numbers are observed in some parts and lower numbers in other parts of the age distribution of immigrants, it is possible that cohort effects may be present. To determine the extent of these effects, the table in Appendix 9.1 offers the distribution of age at immigration for each age from 35 to 55. While there are no parts of the distribution

that are close to zero indicating that immigration happened across all ages and at all points in time relevant to the selected sample here, there is one area of the distribution with notably lower population sizes and one with notably higher sizes.

In addition to being unevenly distributed by age, immigrants are generally also unevenly distributed geographically across Canada. Larger proportions of immigrants normally immigrate to the major cities in Canada. The set of immigrants here should not differ in this respect, as incentives for them to move away from where they are first established as children are few. So, they are expected to be largely concentrated in major cities and in the larger provinces. To illustrate this, Table 3 shows the geographical distribution of the immigrant sample between the 10 provinces and what portion of them live in a Census Metropolitan Area (CMA).[11]

Table 3: Geographical distribution (province of residence and in CMA)

	Sample Count			Population Total		
Province	**All**	**Females**	**Males**	**All**	**Females**	**Males**
NL	232	110	122	1,163	526	636
PE	146	73	73	752	373	379
NS	1,358	724	634	6,539	3,469	3,070
NB	985	534	451	4,494	2,478	2,016
QC	16,191	8,300	7,891	82,363	42,258	40,105
ON	76,216	39,762	36,454	389,901	203,011	186,890
MB	3,758	1,870	1,888	19,523	9,748	9,775
SK	1,100	571	529	5,438	2,842	2,596
AB	12,071	5,996	6,075	61,985	30,663	31,322
BC	23,956	12,236	11,720	122,652	62,822	59,829
Total	**136,013**	**70,176**	**65,837**	**694,810**	**358,190**	**336,619**
not in CMA	18,580	9,617	8,963	89,753	46,405	43,348
in CMA	117,433	60,559	56,874	605,057	311,785	293,271

[11] A Census Metropolitan Area (CMA) is an area consisting of one (or in some cases several) municipalities situated around an urban core with a total population of at least 100,000 where more than 50,000 living in the urban core (Statistics Canada, 2007).

Looking across the provinces, over 90% appear to live in the four biggest provinces with well over half of them living in Ontario alone. Similarly, over 85% overall (from all provinces) appear to live in a CMA.

These geographical characteristics are important to note as they too can influence employment income outcomes. Working in a big city can mean having a higher employment income only because of differences in the cost of living. For this reason, the analysis that follows must in some way take account of these differences to ensure that they are not responsible for any part of the final effect. Consequently, the outcome variable (employment income) is adjusted, through a regression, for differences in age and geographical residence prior to being used in the analysis of discontinuities.

4.4 Outcome variable: employment income

As has already been mentioned, the outcome used in this analysis is employment income. Based on the census definitions, employment income (Empin) includes wages and salaries, and self-employment income (that is, net farm income, or net non-farm income from unincorporated business or professional practice) (Statistics Canada, 2007). This variable is of course only available and collected for the non-institutionalized population aged 15 years or over at the time of the census. The sample of immigrants naturally falls into this category by the nature in which it was selected, so all individuals have a value for their employment income.

Changes made to the 2006 Census pertaining to this outcome variable are important to note here. For the first time respondents were given the option of granting Statistics Canada permission to obtain information on their income directly through their tax records. The introduction of this new mode of collection was meant to reduce respondent

burden and improve data quality for all income variables collected by the census. Indeed, the response to the tax option was very positive, with 82.4% of the eligible population agreeing to it (Statistics Canada, Income and Earnings Reference Guide, 2008). As a result, the precision of all income variables is higher and to some degree more accurate. With higher accuracy also comes greater variability for the income variables since the values will not tend to be as rounded as they would be if they were self-reported.

4.5 Countries of origin and language distance mapping

Drawing on the work by Chiswick and Miller (2005) related to linguistic distance, the analysis in this paper also assesses whether discontinuities in the employment income of immigrants (at points related to the CPH) differ between groups of immigrants coming from countries where the spoken language is far from English and others where it is closer to English. Chiswick and Miller find that immigrants having mother tongues more distant from English have a harder time becoming proficient in English. Consequently, discontinuity effects may be different for immigrants from such different linguistic backgrounds.

Taking advantage of the mapping presented by Chiswick and Miller (2005) of the correspondence between languages and linguistic scores, a similar map is constructed for the languages spoken by the immigrants in the Canadian Census. However, for the purposes of this analysis, the interest is not so much in assigning a linguistic score to the particular languages spoken by the selected set of immigrants, but more in assigning a linguistic score to the country of birth of the immigrants. Because the mother tongue question on the Census refers to the language "first learned and still understood" (Statistics Canada, 2007), some of the immigrants in the selected sample may have

responded with English as their mother tongue when in reality they were first exposed to a different language which they no longer understand. So, in order to avoid having assimilation effects mixed into the analysis, a language score is associated to each immigrant's country of origin[12].

First, using Chiswick and Miller's (2005) table[13] of language to language score correspondence, a mapping was created between the language codes listed in it and the corresponding language codes found in the Canadian Census. While both direct codes and close codes were used for this mapping, not all of the languages from Chiswick and Miller's table were found in the Canadian Census language code set since their codes are from the US Census. The result of this mapping is shown in Table 4 with the languages that did not match (that is, those that were not found on the Canadian Census) highlighted in grey. Once this mapping was done, the language scores were classified into three categories: Low (L), Medium (M), and High (H). The Low category consisted of languages that are more distant from English, having a language score lower than 2.00. The Medium category consisted of languages that are of medium distance to English, having a language score equal to 2.00. And, the High category contained all the languages which are close to English, having linguistic scores of more than 2.00. These categories are shown in the last column of Table 4.

[12] This approach is chosen for the analysis that follows; however, it can have potential shortcomings. It is possible that there is some measurement error associated to this method of classifying the sample of immigrants into language categories using their place of origin rather than the direct language they report as their mother tongue. This can occur due to the possibility that someone from a non-English speaking country could indeed have learnt English as the first language. However, given how broad are the language score categories, the approach taken here seems adequate.

[13] A copy of this table is available in Appendix 9.2.

Table 4: Language Score categories mapping

Chiswick's language(s)	Chiswick (2005) Language score	Canadian Census (2006) Language codes	Language score Category
Afrikaans	3.00	none	H
Danish	2.25	6	H
Dutch, Flemish, Frisian	2.75	3, 4, 5	H
French, Provencal, Patois, Creole, Haitian, Cajun	2.50	2, 123	F
German, Austrian, Swiss, Pennsylvania Dutch, Yiddish, Luxembourgian	2.25	10, 11, 12	H
Italian	2.50	16	H
Norwegian, Icelandic, Faroese	3.00	7, 8	H
Portuguese, Papia Mentae	2.50	13	H
Romanian, Rumanian, Rhaeto-Romanic, Romansch	3.00	15	H
Spanish, Catalonian, Ladino	2.25	14	H
Swedish	3.00	9	H
Indonesian, Bugginese, Moluccan, Achinese, Balinese, Cham, Javanese, Madurese	2.00	none	M
Malay, Bahasa	2.75	67	H
Swahili, Bantu, Bembe, Kikuyu, Kinyarwanda, Luganda, Ndebele, Shona, Tonga, Xhosa, Zulu	2.75	83, 84, 146, 147	H
Amharic, Tigrigna	2.00	75, 77	M
Bengali	1.75	48	L
Bulgarian, Macedonian	2.00	27, 33	M
Burmese	1.75	none	L
Czech	2.00	32	M
Dari → Tadzhik	2.00	none	M
Farsi, Dari, Persian, Pushto, Pashto Afghani, Kurdish, Balochi, Ossete	2.00	38, 39, 49	M
Finnish, Estonian	2.00	80, 81	M
Greek	1.75	18	L
Hebrew	2.00	72	M
Hindi, Sanskrit, Asian Indian, Punjabi, Panjabi, Marathi, Konkani, Gujarathi, Bihari, Rajasthani, Bhili, Romany	1.75	46, 40, 42, 43, 44	L
Hungarian	2.00	82	M
Lao → Thai	**1.50**	**59 --> 58**	**L**
Cambodian, Khmer	2.00	63	M
Mongolian, Tungus, Tibetan	2.00	137	M
Nepali	1.75	none	L
Polish, Slovak, Windish, Lusatian	2.00	34, 36	M
Russian, Bielorussian, Ukranian	2.25	25, 26, 35	H
Serbocroatian, Bosnian, Slavic, Yugoslav, Croatian, Serbian, Slovene	2.00	28, 29, 30, 31, 37, 128	M
Sinhala, Maldivian	1.75	45	L
Tagalog, Filipino, Minangkabau, Sundanese, Bisayan, Ilongo, Visayan, Sebuano, Cebuano, Pangasinan, Ilocano, Igorot, Bikol, Pampangan,	2.00	68, 138, 70, 139	M

Chiswick's language(s)	Chiswick (2005) Language score	Canadian Census (2006) Language codes	Language score Category
Gorontalo			
Thai, Burmese, Karen, Kachin, Laotian	**2.00**	**58, 59?**	**M**
Turkish, Uzbek, Uighur, Azerbaijani, Turkmen, Yakut	2.00	78, 79, 143	M
Vietnamese, Muong	1.50	62	L
Arabic, Syriac	1.50	71	L
Mandarin, Fuchow, Formosan, Fukien, Hokkien, Min Nan, Taiwanese, Wu, Shanghainese	1.50	64, 134, 135, 136	L
Japanese, Ainu	1.00	56	L
Korean	1.00	57	L
Cantonese, Chinese, Min, Hakka, Kan, Hsiang, Toishan, Mien, Miao, Hmong	1.25	60, 65, 66	L

Note: The language category "F" is defined by French and languages similar to it, such as Creole. This is considered as a separate category because French is one of the official languages of Canada. Similarly, though not listed in this table, a language category is also created for English (E).

In the process of creating this language to language score mapping, some discrepancies were discovered in Chiswick and Miller's table (highlighted in bold above). For example, Burmese appeared in two separate sections with a different language score assigned to it in each instance. Fortunately, Burmese was not a language that appeared in the Canadian Census code set, so this turned out not to be problematic. However, in the case of Thai and Laotian, the discrepancy was important. In one instance, Laotian was listed as the primary language having a language score of 1.50, with Thai listed as being in the same group. And, in the second instance, the reverse is listed (Thai being primary, and Laotian being in the same group) but with a language score of 2.00. With the language scores being so different, the language group assigned to these languages would change depending on which score was taken as being "correct". To handle this issue, the two languages were considered to be independent with the primary language having the "correct" score assigned to it. So, Laotian was given a score of 1.50 putting it in the Low category, while Thai was given a score of 2.00 putting it in the Medium category.

Once language score categories were defined for each language wherever possible, corresponding groupings of countries of origin were constructed for each of the linguistic score categories. To do this, the full sample of immigrants from the Census data (not just the immigrant children and young adults, who are 35 to 55 years old) was used together with Table 4, to assign a language score category for each reported mother tongue. Then, for each country of origin, the distribution of the language score categories was computed and a 75% cut-off point[14] was used to determine which country of origin should be assigned which language score category. For example, if over 75% of the people coming from a particular country had a language score category of "L", then it was considered that in general people coming from that country speak a language that has a low linguistic score and is therefore distant from English.

In addition to the Low, Medium, and High categories, a French and English category was also created in the process. As a result, over 90% of the immigrants in the selected sample were assigned a language category using this approach. A detailed list of the countries which were mapped to each language category can be found in the tables of Appendix 9.3.

[14] This cut-off point was also conditional on there being more than ten individuals representing the language category in the sample from that country. This sample size constraint was imposed so that strange results would be minimized (for example, a family from one origin moved to some small country for a few years where the children were born and the language is different, then moved to Canada).

5 Analytical Methods

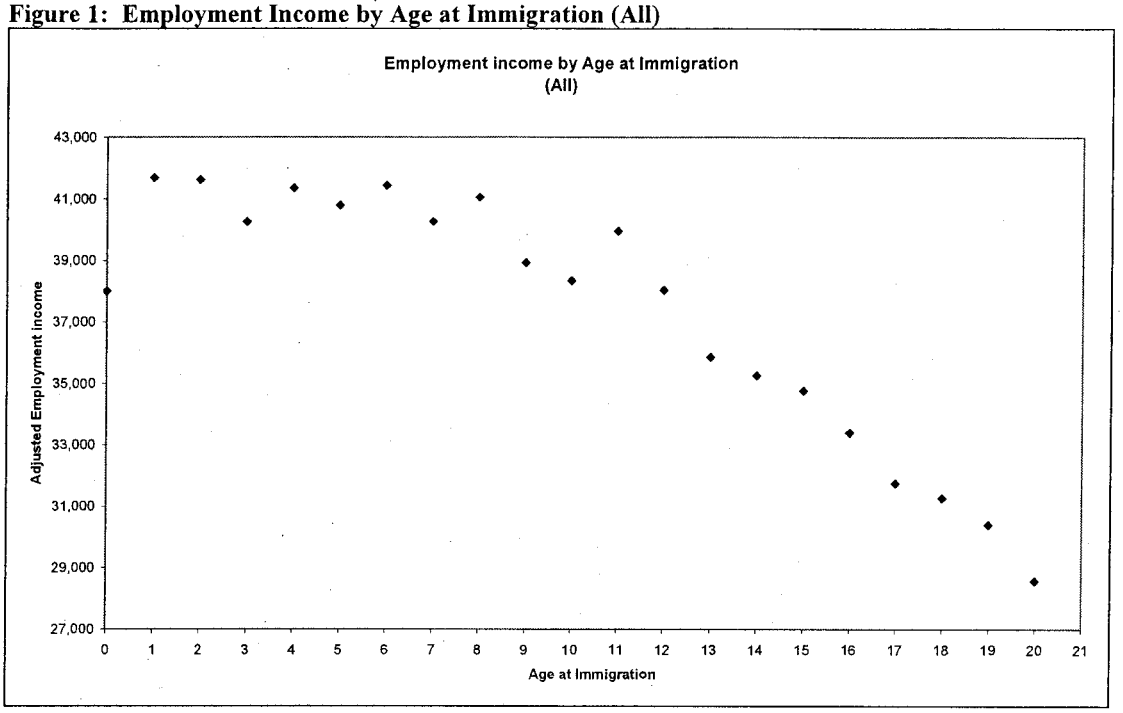
5.1 Regression Design for capturing discontinuities in employment income

The CPH is not an intervention in the sense that there is no policy or rule that formally defines a cut-off point or dictates how to define its "treatment" in a regression framework. However, despite the debate around its existence, discontinuities in the employment income function at points that coincide with the CPH can nevertheless be suggestive. To capture these effects, elements from a sharp RD design are used to define the model. Thus, as a first step, to validate the incorporation of RD components in the model, the degree to which the CPH fits the assumptions of the RD design is evaluated, specifically with respect to the presence of the required relationships between the involved variables.

The practical guide of Imbens and Lemieux (2008) provides a convenient and comprehensive approach for testing these assumptions through graphical analysis. Following their strategy, the first plots of interest are displayed in the three figures below. The graphs show the average (adjusted) employment income for each age at immigration for the full sample of immigrants, the sample of male immigrants, and the sample of female immigrants. Focusing on the graph of the full sample, the employment income series appear to be continuous (with some variation in the early years) until the age of 12. At the age of 12 there seems to be a small break as the series jumps down between the ages of 12 and 13 and then looks to be continuous onwards with less variation than prior to the break. This jump at the age of 12 coincides with the CPH. Also, it looks to be the only clearly observable break in the graph, as well as the point separating observations of

higher variability from observations with noticeably lower variability. So, the existence of discontinuities at other points is considered unlikely.

Figure 1: Employment Income by Age at Immigration (All)



When a distinction is made between the two genders, a somewhat different picture emerges. The effect seen in the graph for the full population is again present in the graph for males. However, in this graph, it is not too clear whether the break is at the age of 12 or the age of 11. While the gap between the age 11 and the age 12 seems to be just as large as the one between the ages of 12 and 13, the jump down from 12 to 13 still stands out as a more distinct break between the observations, again separating the higher variability portion in the series from the lower variability one. Due to the higher variation observed in the earlier years, this will be considered as the only point of discontinuity, and despite this variation, the relationship between the outcome (employment income)

and the forcing variable (age at immigration) looks to be relatively continuous on either side of the age of 12.

In contrast, for females there is some evidence that a discontinuity may exist at the age of 12 or at the age of 15. If the variation in the observations is considered again, the break at the age of 15 seems to be more distinct and even possibly larger in magnitude. This was not at all the case for males, but nevertheless, the gap at that the age of 15 is a point that should certainly be considered as a possible discontinuity for females.

With these two pictures in mind, it is apparent that any formal tests of the existence of such discontinuities must be done separately for males and females due to the different critical ages that may be affecting for the two genders.

Figure 2: Employment Income by Age at Immigration (Males)

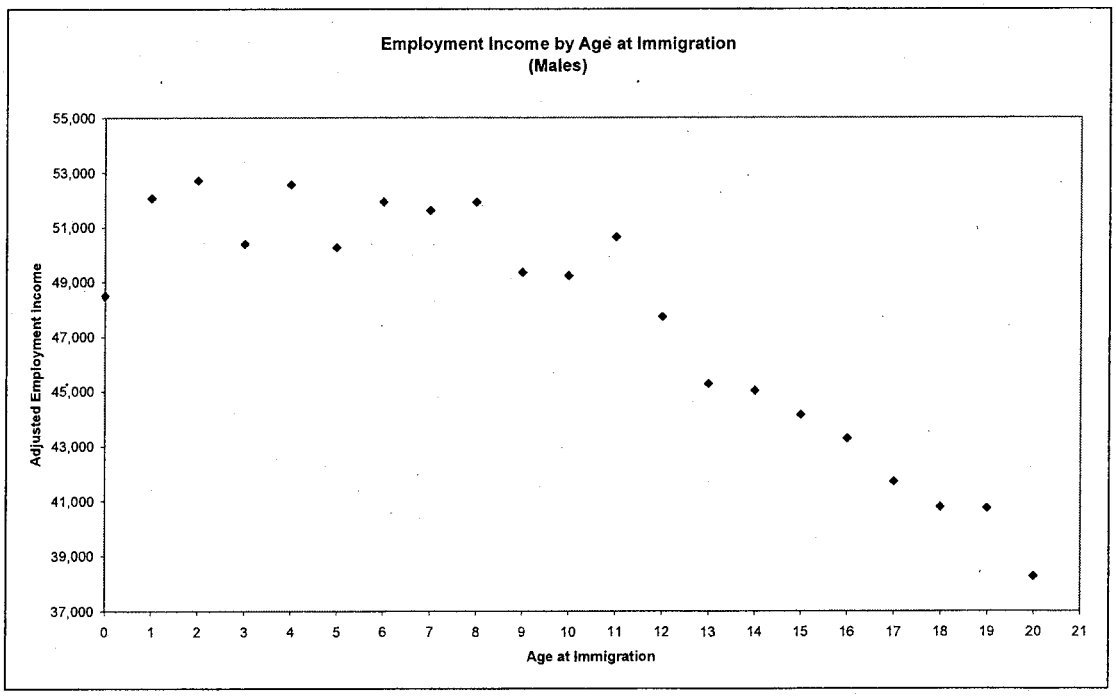
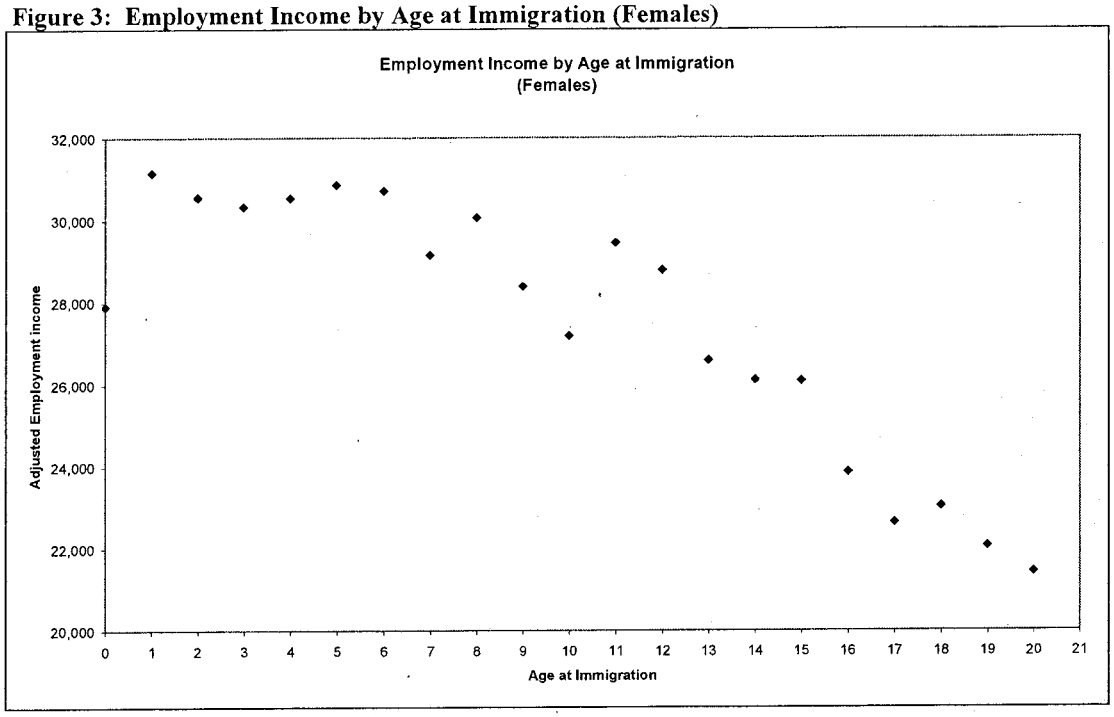


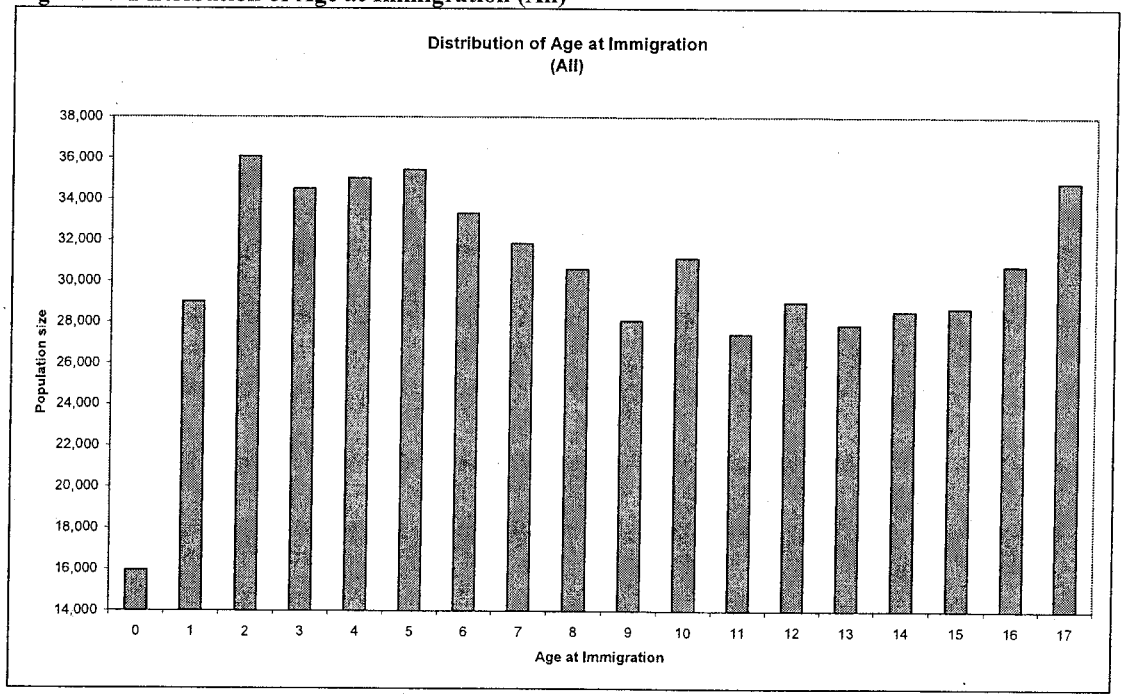
Figure 3: Employment Income by Age at Immigration (Females)



For the verification of the third assumption, the distribution of the forcing variable (age at immigration) is shown in Figure 4. The ages of 18, 19, and 20 have been removed from the graph as these individuals could be immigrating on their own and this could distort the distribution. To satisfy the condition that the observations on either side of the critical point are otherwise identical, the graph should be as close as possible to a uniform distribution. If that is the case, then it can be assumed that individuals are not manipulating whether or not they fall in the treatment category. In the context of the CPH, this assumption translates to mean that when parents immigrate, they are not aware of the CPH, and, therefore, their decision of when to immigrate is not based on the age of their children. So, if there is no reason for parents to specifically bring their children when they have reached a particular age, the distribution of age at immigration in the sample of immigrants here should be more or less uniform.

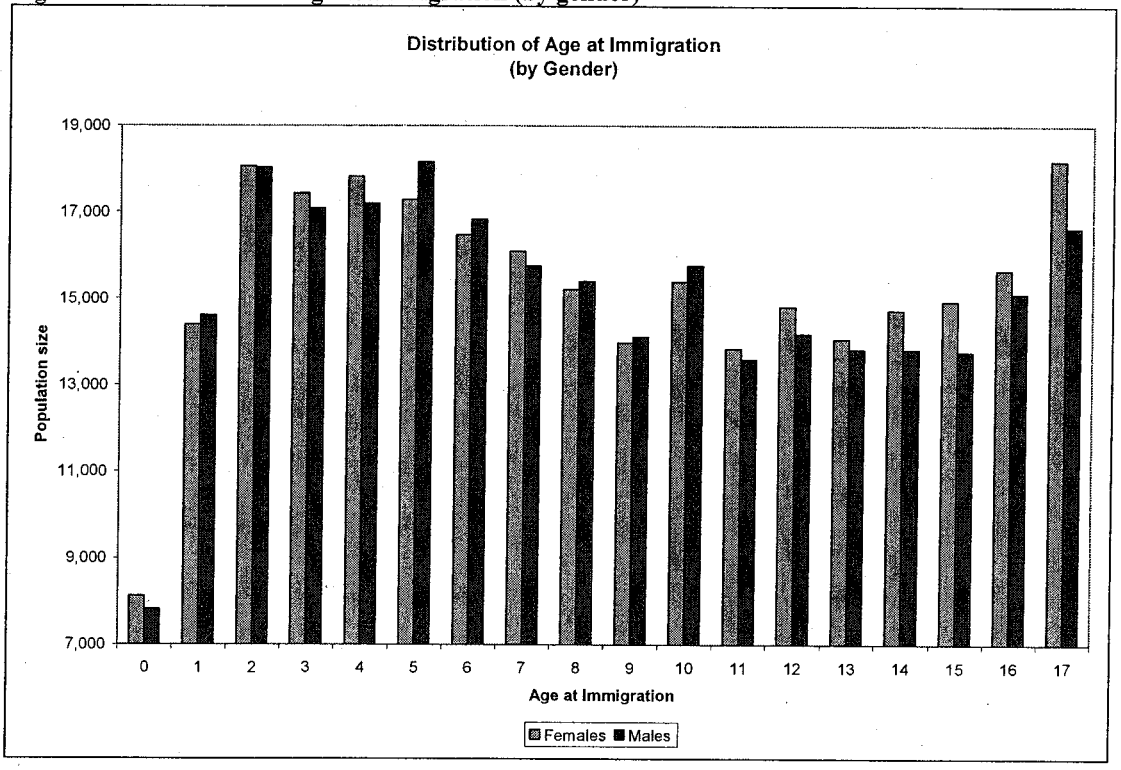
Looking at Figure 4, the required uniformity does not appear to be as present as might be desirable. Immigrating as a small baby (age of zero) seems to be a lot less common than immigrating at other ages, and there appear to be less people immigrating in the ages of 11 to 15. One possible explanation for the lower numbers of baby immigrants may be that the zero category is reflecting part of the overestimation error in the age at immigration variable discussed earlier. As for the smaller numbers observed for the ages of 11 to 15, it may be related to the fact some parents, who invest more in their children, may be in some way timing their immigration decision by somehow taking into account what impact it might have on their children. Nevertheless, the distribution is not all that far away from uniform hovering a bit above or a bit below 32,000 for each age at immigration.

Figure 4: Distribution of Age at Immigration (All)



The distribution of age at immigration is shown by gender in Figure 5. The pattern is the same as the one observed for the overall population, and the proportion of males and females seems to be roughly the same, except for some small differences appearing among those who arrived at an age of 14 or older. For this group of immigrants, there are slightly more females than males for every age at immigration after 14.

Figure 5: Distribution of Age at Immigration (by gender)



The fourth assumption is related to the smoothness of the other covariates of the outcome with respect to the forcing variable. Satisfying this assumption ensures that the observed discontinuity is exclusively a result of the change in treatment as defined by the forcing variable. To account for this, the outcome is adjusted, through a regression, such that all differences from a set of covariates are netted out.

5.2 Appropriate Estimation

As mentioned previously, as a preliminary step to the analytical model, the employment income variable was adjusted (through a regression) for differences within the immigrant population of interest related to place of residence, and age[15]. The adjustment was done by running a regression of the employment income variable on a set of variables denoting place of residence and age. Then the resulting residuals were added to the estimate of the intercept to create the adjusted employment income variable. By doing this, all existing differences (in this set of variables) for the immigrant population of interest were "netted out" ahead of time. As a result, differences in these variables cannot be responsible for the effects observed in the regression models in the end. Equation 3 gives the specification for this preliminary regression.

Equation 3: Regression for Adjusting Earnings

$$Empin_i = \beta_0 + \beta_{NL}NL_i + \beta_{PE}PE_i + \beta_{NS}NS_i + \beta_{NB}NB_i + \beta_{QC}QC_i + \beta_{ON}ON_i + \beta_{MB}MB_i + \beta_{SK}SK_i + \beta_{AB}AB_i + \beta_{BC}BC_i + \beta_{cma}INCMA_i + \beta_{age}agedev_i + \beta_{age2}agedev2_i + \varepsilon_i$$

In this equation, each provincial variable is a dummy variable indicating whether or not individual *i* lives in that province. Similarly, the INCMA variable is a dummy variable indicating whether or not the individual lives in a Census Metropolitan Area. The variable "agedev" represents the individual's age as a deviation from the mean age of the sample of immigrants (those aged 35 to 55 who arrived at or before the age of 20), and "agedev2" is the square of that variable. Furthermore, rather than excluding one of the provincial dummy variables, making the excluded province the reference province,

[15] This adjustment is done to correct for life-cycle differences, so that such differences would not influence the end results. As an alternative approach, it would also be possible to include the variables used for this adjustment directly into the estimation model for employment income discontinuities. However, because the model used for the estimation is a model using group-means, it was more convenient to do the adjustment separately.

the regression is instead restricted such that the provincial coefficients all add to zero. With this restriction it is the "average province" that is the reference point in the regression instead of a particular province, as is commonly done. The regression is also weighted by the composite weight attached to the Census file so that the full population of interest is represented appropriately. The results of the regression are shown in Appendix 9.4.

With the individual employment income adjusted in this way, the a model is implemented following much of the methodology from Lemieux and Milligan (2008), who use a RD model in their study of social assistance using the Canadian census. The approach of Lemieux and Milligan (2008) is practical because the construction of their RD model is simple. They work from the following base RD model.

Equation 4: RD Base Model

$$Y_{ia} = \beta_0 + \beta_1 TREAT_{ia} + \delta(a) + \varepsilon_{ia}$$

In this equation, Y_{ia} is the outcome variable, in this case – (adjusted) employment income, for individual i of age a, and the effect of forcing variable on the outcome variable is captured by the function $\delta(a)$, which Lemieux and Milligan (2008) emphasize must be smooth (in other words, that differences in the outcome variable are the only source of discontinuity around the critical point). $TREAT_{ia}$ is a treatment dummy variable defining the critical point at which the discontinuity happens. However, since the forcing variable in their study (age) is not a continuous variable (it is a categorical variable), Lemieux and Milligan (2008) construct their model using the cell means for each value of the forcing variable. Furthermore, because the forcing variable in their study is age, which is measured at the time of Census collection, but the outcome variable

is the employment rate, which is measured as of the end of the previous year, Lemieux and Milligan use fuzzy RD for their analysis. As a result, the $TREAT_a$ variable in their model is not a binary variable representing the treatment group; rather, it reflects the expected value of the treatment around the discontinuity point.

Equation 5: RD Cell-Means model

$$Y_a = \beta_0 + \beta_1 TREAT_a + \delta(a) + \varepsilon_a$$

With this specification, Lemieux and Milligan test the fit of their model using the Goodness of Fit (GOF) statistic: $GOF = \sum_a \left(\frac{\hat{e}_a^2}{V_a} \right)$ where V_a is the sampling variance of the cell mean Y_a. The GOF then follows a Chi-square distribution with $N - k$ degrees of freedom.

The strategy used in this paper mimics the sharp version of the cell means RD specification outlined by Lemieux and Milligan. With this implementation discontinuities in the employment income function around points associated to the CPH are explored for various groups of the selected immigrant population. The variable Y_a is the average (adjusted) employment income for age at immigration a, and the $TREAT_a$ variable is the treatment. For the models where the critical age of interest is 12 (that is, for the overall sample and the male sample of immigrants), it is defined as $TREAT_a = \begin{cases} 0 \text{ if } a \leq 12 \\ 1 \text{ if } a > 12 \end{cases}$. And, for the models where the critical age of interest is 15 (that is, for the female sample of immigrants), it is defined as $TREAT_a = \begin{cases} 0 \text{ if } a \leq 15 \\ 1 \text{ if } a > 15 \end{cases}$. The analysis then consists in estimating the effect, β_1, of the treatment on the outcome variable. If β_1 is statistically

significant, then it gives a measure of the discontinuity at the critical age in question. It can then be interpreted that the discontinuity that is observed results from the change in treatment at that point (in other words, that the employment income of immigrants who came after the age of 12/15 are higher/lower precisely because they arrived after that age rather than before).

The $\delta(a)$ function in the model is a function of the forcing variable (age at immigration) capturing its effect on the outcome variable. In Section 6, results are shown for five different forms of the $\delta(a)$ function in order to examine which specification is the most appropriate for the relationship between age at immigration and employment income. The function takes on the following forms: linear in age, age-squared, quadratic in age, age-cubed, and cubic in age.

6 Analysis and Results

6.1 Regression results – overall and by gender

The results of the regression estimates are presented in this section. The analysis begins by considering the model for the full population of the immigrant sample of adults, aged between 35 and 55, who arrived in Canada at or before the age of 20, and then moves on to look at how the effects differ by gender. As was observed in the graphs of the average (adjusted) employment income against the age at immigration (Figure 1 – Figure 3), the hypothesized discontinuity was at the age of 12 for the overall population and for the male population, but for the female population, it occurred at the age of 15. The regressions are, therefore, designed and specified to capture the effects of these critical ages, 12 and 15 respectively. One-sided T-tests are used to determine the level of significance of the treatment effect because the analysis here is motivated by the CPH which suggests that observed discontinuities should occur in one particular direction. For the first set of models, all available observations are used, and all five specifications for the functional form of the $\delta(a)$ function are tested. The results for each form of the model are presented in Table 5 for the overall sample, as well as separately by gender.

In the first panel of the table, for the total population, three of the five models show to have a significant treatment effect (for the critical age of 12) at least at the 5% level of significance: the linear model, the quadratic model, and the age-cubed model. The treatment effect for two of these models is significant even at the 1% level. In the linear case, the results suggest that those who immigrated as children after the age of 12 make on average over $3,700 per year less than those who immigrated at the age of 12 or

before. This effect is smaller for the other two models, but nevertheless, even the effect seen in the age-cubed specification (– $2,100) is still significant.

For the critical age of 15 the linear model shows to be significant once again for the overall population, at the 1% level of significance, with a stronger negative effect of over $3,900. To better understand these discontinuities and what is happening at these two critical ages, the models are re-fitted for males and females separately, so as to distinguish between the different critical ages for each gender (12 for males and 15 for females).

Table 5: Regression results overall and by gender

	Model	Critical age = 12			Critical age = 15		
		Treatment	Adjusted R-squared	GOF	Treatment	Adjusted R-squared	GOF
All	Linear in age	-3,728 **	0.89	1%	-3,965 **	0.91	1%
	Age-squared	-1,405	0.96	1%	-1,113	0.96	1%
	Quadratic in age	-1,422 *	0.97	5%	-24	0.97	1%
	Age-cubed	-2,102 **	0.97	1%	6	0.96	1%
	Cubic in age	-1,082	0.97	5%	-199	0.97	5%
Females	Linear in age	-2,539 *	0.89	1%	-3,277 **	0.93	1%
	Age-squared	-1,012	0.94	1%	-1,532 *	0.95	5%
	Quadratic in age	-1,040	0.94	1%	-1,382	0.95	5%
	Age-cubed	-1,729 *	0.94	1%	-1,153	0.93	1%
	Cubic in age	-528	0.94	1%	-1,545	0.95	5%
Males	Linear in age	-4,597 **	0.90	1%	-3,762 **	0.89	1%
	Age-squared	-2,375 *	0.95		-545	0.94	5%
	Quadratic in age	-2,350 *	0.96		661	0.94	
	Age-cubed	-3,060 **	0.96		702	0.93	1%
	Cubic in age	-1,873	0.96		349	0.95	

Notes:

* Significant at the 5% level (one-sided T-test)

** Significant at the 1% level (one-sided T-test)

The GOF column reports the level of significance for the Goodness of Fit Test when it is significant at the 5% or 1% level.

As was expected from the graphical analysis, for females, some of the models come out with significant treatment effects for the critical age of 12, and some have significant

treatment effects for the age of 15. However, for the critical age of 15, the linear model seems to have the only instance of the treatment effect being significant at the 1% level, and this is also the one with the largest negative effect of over $3,200. This represents 12.2% of the overall average (adjusted) employment income for females selected in this study.

In the case of males, four out of the five models result in significant treatment effects for the critical age of 12. Once again, the linear model has the largest negative treatment effect of close to $4,600, corresponding to 9.7% of the overall average (adjusted) employment income for these males. The treatment effect in the other models is smaller than this, but it seems to be of similar magnitude. Further, with all the models having such a good fit (the adjusted R-squared is high, and the Goodness of Fit is significant in the linear case), it indicates that the results are robust in so far as that the functional form used in the specification does not affect the results to a great degree (that is, the treatment is significant irrespective of the functional form).

These results reflect the important differences between males and females, suggesting that discontinuities at points similar to those defining the CPH are indeed detectable, but are of different magnitudes and at different "critical" ages for the two genders. However, to be able to make such conclusions with more certainty, the model should be refined a little to include a few other elements of the theoretical RD design, which is what drives the validity of exploring discontinuities in this way. The following section develops this idea and discusses the corresponding results.

6.2 Regression results – refining the window

One of the fundamental characteristics of the RD design is that the analysis is done in some neighbourhood of the critical point. More specifically, Imbens and Lemieux (2008) recommend using local linear regression with a set number of observations on either side of the discontinuity point. The intuition for this is that an uneven set of observations on either side of the critical point could have a dramatic impact on the treatment effect. Any extra observations on one side can greatly skew the final estimates in a particular direction, whereas keeping the same number of observations around the critical point ensures that the "treatment" group and the "control" group are comparable and of similar sizes.

Therefore, in this section of the analysis, the linear model is redefined to incorporate this component by including different windows of observations on either side of the critical points of 12 and 15. The reason for choosing to continue with the linear model in this next step is twofold. In part it is owing to the fact that for this simple model specification, the initial results shown in Table 5 were the strongest, but secondly to benefit from another feature of the RD design which permits for a more careful examination of what happens around the discontinuity point. Thus, following the advice of Imbens and Lemieux (2008), the linear regression function is fitted only for the observations which are within a certain distance of the critical point. In this way, results would only depend on observations close to the critical point, and not ones which are far away from the discontinuity and can, therefore, have no pertinence to it.

For the critical age of 12, the largest possible window that can be used is one of eight observations on either side if the critical point. For the age of 15, however, at most

five observations are available on each side. Table 6 displays the results for the RD model specified for each possible window around the critical points of 12 and 15.

Table 6: Regression results for different windows

Linear Model		All			Males			Females		
		Treatment	Adj-R^2	GOF	Treatment	Adj-R^2	GOF	Treatment	Adj-R^2	GOF
Critical age = 12										
Window	8	-1,190	0.95	1%	-2,351 *	0.94	5%	-697	0.93	1%
	7	-1,281	0.96	5%	-2,118 *	0.97		-760	0.89	1%
	6	-1,362	0.94	5%	-1,786	0.96		-1,013	0.83	1%
	5	-1,199	0.92	5%	-1,825	0.93		-682	0.77	1%
	4	-2,183	0.87		-2,714 *	0.89		-1,787	0.59	1%
	3	-2,457	0.82		-2,410	0.84		-3,056	0.58	
	2	-939	0.90		-827	0.74		-1,620 *	1.00	
Critical age = 15										
Window	5	318	0.98		988	0.95		-1,242	0.96	
	4	-348	0.97		-105	0.94		-1,379	0.93	
	3	-713	0.96		-100	0.95		-2,151 *	0.94	
	2	-210	0.88		402	0.94		-1,562	0.88	

Notes:

* Significant at the 5% level (one-sided T-test)

** Significant at the 1% level (one-sided T-test)

The GOF column reports the level of significance for the Goodness of Fit Test when it is significant at the 5% or 1% level.

Males appear to have significant results for several of the possible windows around their critical point of the age of 12. As the window narrows, the negative treatment effect appears to change a little, but in general, it has similar values for all windows except for the window with 2 observations. For females, although the effects are smaller in magnitude, they are also of similar values for all windows except the window of three observations, which is significant at the 5% level with a negative treatment effect of over $2,100. This can be seen as one more indication of some degree of robustness in the results as the size of the window used in the model does not appear to have much influence on the magnitudes of the treatment effect, only on their significance in some cases.

The interpretation of these numbers gives them more specific meaning. The age of 12 is the age at which several changes occur in a child's life. One important such change is the onset of puberty as argued by the proponents of the CPH. Some of the results presented so far in this study suggest that there exists variation in the age at which puberty starts. The second important change occurring around the age of 12 is the beginning of the transition from elementary school to high school. The age of 15, on the other hand, is the age at which children have usually just recently started high school, which lasts for four years. Given that these changes are important in a child's development, and that the CPH is centered on how learning a second language is affected during this period of time, it seems appropriate to consider the 7 year window for the critical age of 12 and the 4 year window for the critical age of 15 as the defining model for this analysis.

Using this specification, the model not only fits the data and gives significant results for the treatment effect, but also gives the resulting effects a more meaningful interpretation. For males, the results now suggest that if a child immigrates in the seven years after the age of 12 (the teenage years of transition from elementary school to high school and of high school itself), then he would make over $2,100 less as an adult than if he immigrated in the seven years leading up to the age of 12 (during the years of elementary school and before the beginning of puberty). This is an important result, representing 4.5% of average male (adjusted) employment income. For females, the effect is a little stronger. Those who immigrate in the four years after the age of 15 (the high school years) appear to make close to $1,400 less than those who immigrate in the

four years leading up to 15 (the years of transition to high school), representing 5.1% of their average (adjusted) employment income.

6.3 Regression results – by language category

This section examines whether the effects observed so far change when English is learnt from a different base language. In particular, regression results are presented for each language category as defined in Section 4.5, classifying the selected immigrants by the distance (from English) of the language spoken in their country of birth. In this way, the effects can be measured separately for people coming from countries where the spoken language is similar (that is, close) to English, and for people coming from countries where the spoken language is very different (that is, distant) from English. Table 7 gives these results.

There are two interesting results shown in this table. The first is that the Low category, representing immigrants coming from countries where the spoken language is distant from English, has significant negative results for males in the magnitude of over $5,600. This is a substantial difference, corresponding to 13.1% of this group's average (adjusted) employment income. For the overall population in that category, the effect is again significant, but is about half the magnitude of that for males (or 8.4% of the overall average (adjusted) employment income).

Table 7: Regression results by language category

Linear Model		Critical age = 12 Window = 7			Critical age = 15 Window = 4		
	Language Category	Treatment	Adj-R^2	GOF	Treatment	Adj-R^2	GOF
All	English	-1,904 *	0.89		1,686 *	0.91	
	French	50	0.24		-3,468	0.34	
	High	-440	0.90		-635	0.92	
	Medium	303	0.41	1%	-7,818	0.54	1%
	Low	-2,886 *	0.97		674	0.85	
Females	English	-1,006	0.71		1,365	0.68	
	French	2,037	0.06	5%	431	0.48	
	High	23	0.88		-1,842	0.89	
	Medium	-3,419	0.35	1%	-6,306	0.20	1%
	Low	-528	0.92		-2,137	0.93	
Males	English	-3,262 *	0.84		953	0.37	
	French	-1,796	0.10		-9,308	0.19	
	High	-634	0.78		385	0.65	
	Medium	3,553	0.27	1%	-10,005	0.48	5%
	Low	-5,626 *	0.90		2,446	0.51	1%

Notes:

* Significant at the 5% level (one-sided T-test)

** Significant at the 1% level (one-sided T-test)

The GOF column reports the level of significance for the Goodness of Fit Test when it is significant at the 5% or 1% level.

The second point to note is that for both the overall population and for the males-only portion, the English category model captures a significant discontinuity (for the age of 12). The negative treatment effect is as large as $3,200 for males, but when averaged out with the females, it is only around $1,900 (for the overall population). This is a particularly curious result since the expectation is that anyone immigrating to Canada from a country where the spoken language is English or French should not be experiencing such discontinuities in their employment income function at any age at immigration.

This result suggests that there may be other factors influencing the employment income of these immigrants. As such, in an effort to pick out which portion of the sample

is precisely contributing to this overall effect, it motivates further investigation into the breakdown of these effects within some smaller domains. This is what is presented in Table 8, where three of the categories (English, Low, and High) are subdivided into three sub-categories: the two countries which make up the largest proportion of immigrants are separated out from the rest of the countries in each category. For the English category, the USA and the UK become isolated from the remaining English-speaking countries. In the Low category, Hong Kong and India are separated out. And, in the High category, it is Italy and Portugal which make up the largest number of immigrants. The Medium category and the French category are not considered for the smaller domain analysis because separating out specific countries from these categories was not possible due to sample size constraints.

Table 8: Regression results by language category and country

Linear Model			Critical age = 12 Window = 7			Critical age = 15 Window = 4		
	Language Category	**Country**	**Treatment**	**Adj-R^2**	**GOF**	**Treatment**	**Adj-R^2**	**GOF**
All	English	USA	-4,592	0.65		871	0.67	
	English	UK	-3,597 *	0.51		2,212	0.06	
	English	Remaining English category countries	1,729	0.61		1,883	0.20	
	Low	Hong Kong	-2,767	0.70		-28	0.13	
	Low	India	919	0.94		-3,391 **	0.99	
	Low	Remaining Low category countries	-4,529 *	0.93		2,701	0.75	1%
	High	Italy	-448	0.87		781	0.48	
	High	Portugal	-1,779	0.83		2,078	0.89	
	High	Remaining High category countries	538	0.35		-3,036 *	0.90	
Females	English	USA	-890	0.61		-1,096	0.24	
	English	UK	-1,640	0.13	5%	1,798	0.35	
	English	Remaining English category countries	-252	0.35		2,087	-0.17	
	Low	Hong Kong	4,673	0.32		-4,680	0.47	

	Low	India	-2,726	0.87		-4,385 *	0.84	
	Low	Remaining Low category countries	-1,738	0.89		-689	0.83	
	High	Italy	1,062	0.77		-1,849	0.75	
	High	Portugal	1,518	0.84		3,589	0.76	
	High	Remaining High category countries	-1,917	0.12	1%	-4,989	0.42	
Males	English	USA	-8,974	0.38		2,615	0.05	5%
	English	UK	-5,753 *	0.10		916	0.65	
	English	Remaining English category countries	3,690 *	0.40		899	-0.11	
	Low	Hong Kong	-10,240 *	0.34		6,400	-0.24	
	Low	India	2,462	0.77		-2,932	0.87	
	Low	Remaining Low category countries	-6,679 *	0.89		4,475	0.64	5%
	High	Italy	1,014	0.73		2,635	0.12	
	High	Portugal	-6,833 *	0.57		972	-0.35	
	High	Remaining High category countries	2,671	0.30		-1,430	0.65	

Notes:

* Significant at the 5% level (one-sided T-test)

** Significant at the 1% level (one-sided T-test)

The GOF column reports the level of significance for the Goodness of Fit Test when it is significant at the 5% or 1% level.

With the results broken down in this way, some of the questions raised so far can be explained. It is now clear that the strong effects seen in the English category earlier are driven by the part of the population coming from the UK, and particularly the male portion of the UK immigrants. The negative treatment effect among this group shows to be well over $5,700, representing 10.7% of their average (adjusted) employment income. In the USA, on the other hand, no significant effect is observed even if the estimate of the treatment variable shows to be large and negative. For the remaining English category, a significant positive effect is seen for males. Having results with such differing effects between groups of immigrants from different English speaking countries may be another indication that there are other factors at play.

With respect to the High and Low sub-categories, the results show that different countries correspond to different effects even within the same language category. For

example, for Italian males, the effect, though not significant, is positive, whereas for Portuguese males, it is significantly negative. A male immigrant from Portugal arriving in the seven years after the age of 12 would make on average over $6,800, or 16.1% of average (adjusted) employment income, less than a similar male immigrant from Portugal who arrives in the seven years leading up to the age of 12. In the Low category, a similar pattern is seen. There appears to be a positive effect for males coming from India, though again not significant. This result could be partly reflecting the fact that English can be quite common in India and immigrants coming from there may have had exposure to English prior to arriving. However, the effect for males from Hong Kong is highly negative and indeed significant. Male immigrants from Hong Kong who immigrate in the seven years after the age of 12 would make on average $10,240 (19.5%) less than their counterparts who arrived in the seven years before. This is a considerable difference. For the remaining countries in the category of distant languages a negative effect is also present (of almost $6,700, or 14.4% of average (adjusted) employment income).

For females, effects as large as those for males are not seen in the results, but the analysis of detecting a discontinuity was always much weaker as the jumps at the breaking point were smaller and therefore harder to measure. Nevertheless, females coming from India appear to fit the model very well showing a negative treatment effect for the age of 15 of almost $4,400, representing 18.5% of their average (adjusted) employment income.

7 Conclusion

Using the 2006 Canadian Census of Population, this paper investigates discontinuities in the employment income function of adult immigrants who arrived in Canada at a young age. A regression model is developed based on elements of a RD design to identify and measure these discontinuities at points that coincide with the CPH.

The results show that a discontinuity in the employment income function of immigrants, who immigrated at or before the age of 20, is observed at a different point for males than for females. Discontinuities for males appear to be at age at immigration of 12, while for females, they seem to be at age at immigration of 15. The effect is significantly negative for males, putting employment income for those having arrived at ages just after 12 to be over $2,100 lower than that of those having arrived at ages leading up to the age of 12. For females, this effect is close to $1,400, corresponding to a higher proportion of their average (adjusted) employment income.

The results by language category reveal that differences in these effects exist between the two genders within a category, as well as between the categories. For the language category representing the countries with languages distant from English, males were found to have an overall negative effect of $5,600, representing over 13% of their average (adjusted) employment income. Looking at the sub-categories within this group, male immigrants from Hong Kong are found to be the main contributors to this result, have a negative effect of over $10,200, representing 19.5% of their average (adjusted) employment income.

While no effect was detected for immigrants from countries where the spoken language is close to English, a surprising result is discovered for the immigrants coming

from English-speaking countries. A significant (negative) discontinuity of over $3,200 is observed for males in this group. The more detailed results for this category show that immigrants from the USA, and specifically those from the UK, are the ones most affected. This result may suggest that some other factors may be playing a role and may be influencing these discontinuities.

For females, discontinuities are not detected as often as for males due to the fact that the jumps in the employment income functions for females are smaller and therefore harder to identify as statistically significant. However, for females from India, a discontinuity of almost $4,400 was indeed detected at the age of 15. This is an important (negative) effect as it represents 18.5% of their average (adjusted) employment income.

Even though this paper does not explore this, it would be of interest to investigate the measurement error in the age at immigration variable. Of particular concern would be to establish to what degree it may be influencing these discontinuity points. Perhaps in the absence of this error, the discontinuity would be more profound; or, on the contrary, the employment income function may be smoother suggesting no discontinuity exists. Furthermore, as an extension to this work, the effects of age at immigration on other outcomes, such as labour force participation, employment, education levels, and marriage rates, could also be considered. Collectively, this would give a fuller idea of the overall effect of immigrating at different ages.

Given that some results suggest other factors may be responsible for the observed effects, it would be valuable to consider implementing an identification strategy similar to that of Bleakley and Chin (2004) for this data, while maintaining that the effects can differ between the language categories and the two genders. This would specify the

relationship more precisely in the context of causal inference and allow for a more direct interpretation of the resulting effects.

8 References

Angrist, J. and A. Krueger (1999). "Empirical Strategies in Labour Economics." In *Handbook of Labour Economics, 3A*, ed. Orley Ashenfelter and David Card, 1277-1366. Amsterdam: North-Holland.

Angrist, J. and A. Krueger (2001). "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *The Journal of Economic Perspectives,* 15 (4), 69-85.

Angrist, J. and J.-S. Pishke (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: Princeton University Press.

Angrist, J. and V. Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics,* 114 (2), 533-575.

Aydemir, A. and M. Skuterud (2005). "Explaining the Deteriorating Entry Earnings of Canada's Immigrant Cohorts, 1966 – 2000." *The Canadian Journal of Economics,* 38 (2), 641-671.

Baker, M. and D. Benjamin (1994). "The Performance of Immigrants in the Canadian Labour Market." *Journal of Labour Economics,* 12 (3), 369-405.

Bleakley, H. and A. Chin (2004). "Language Skills and Earnings: Evidence from Childhood Immigrants." *The Review of Economics and Statistics,* 86 (2), 481-496.

Bloom, D., G. Grenier, and M. Gunderson (1995). "The Changing Labour Market Position of Canadian Immigrants." *Canadian Journal of Economics,* 28 (4), 987-1005.

Casey, T. and C. Dustmann (2008). "Intergenerational Transmission of Language Capital and Economic Outcomes." *The Journal of Human Resources,* 43 (3), 660-687.

Chiswick, B. and P. Miller (2001). "A Model of destination language acquisition: Application to male immigrants in Canada." *Demography,* 38 (3), 391-409.

Chiswick, B. and P. Miller (2005). "Linguistic distance: A quantitative measure of the distance between English and other languages." *Journal of Multilingual and Multicultural Development,* 26 (1), 1-16.

Chiswick, B. and P. Miller (2008). "A test of the critical period hypothesis for language learning." *Journal of Multilingual and Multicultural Development,* 29 (1), 16-29.

Chiswick, Barry (2007). "The Economics of Language for Immigrants: An Introduction and Overview." IZA Annual Migration Meeting, No. IV, Julian Simon Lectures Series, May.

Corak, Miles (2008). "Immigration in the Long Run: The Education and Earnings Mobility of Second-Generation Canadians." *IRPP Choices*, 14 (13).

Ferrer, A. and W. Riddell (2008). "Education, Credentials, and Immigrant Earnings." *Canadian Journal of Economics*, 41 (1), 186-216.

Ferrer, A., D. Green, and W. Riddell (2006). "The Effect of Literacy on Immigrant Earnings." *Journal of Human Resources*, 41 (2), 380-410.

Hakuta, K., E. Bialystok, and E. Wiley (2003). "Critical Evidence: A Test of the Critical-Period Hypothesis for Second-Language Acquisition." *Psychological Science,* 14 (1), 31-38.

Imbens, G. and T. Lemieux (2008). "Regression discontinuity designs: A guide to practice." *Journal of Econometrics,* 142 (2), 615-635.

Lemieux, T. and K. Milligan (2008). "Incentive Effects of Social Assistance: A regression Discontinuity Approach." *Journal of Econometrics,* 142 (2), 807-828.

Nichols, Austin (2007). "Causal Inference with Observational Data." *The Stata Journal,* 7 (4), 507-541.

Penfield, W. and L. Roberts (1959). Speech and Brain Mechanisms. Princeton: Princeton University Press.

Schaafsma, J. and A. Sweetman (2001). "Immigrant Earnings: Age at Immigration Matters." *The Canadian Journal of Economics,* 34 (4), 1066-1099.

Statistics Canada (2007). *2006 Census Dictionary.* Statistics Canada Catalogue no. 92-566-XWE. Ottawa. December 11. http://www12.statcan.ca/english/census06/reference/dictionary/index.cfm (accessed March 12, 2009).

Statistics Canada (2008). "Income and Earnings Reference Guide, 2006 Census." Statistics Canada Catalogue no. 97-563-GWE2006003. Ottawa. December 4. http://www12.statcan.ca/english/census06/reference/reportsandguides/income-earnings.cfm (accessed December 12, 2008).

Statistics Canada (2008). "Place of Birth, Generation Status, Citizenship and Immigration Reference Guide, 2006 Census." Statistics Canada Catalogue no. 97-557-GWE2006003. Ottawa. July 29. http://www12.statcan.ca/english/census06/reference/reportsandguides/immigration.cfm (accessed November 14, 2008).

US Bureau of the Census (2003). "2000 United States Census of Population and Housing." Technical Documentation, Washington, D. C.

9 Appendices

9.1 Distribution of Age by Age at Immigration

Table 9: Distribution of Age by Age at Immigration

		Sample Count								Population Total						
Age		35	36	37	38	39	40	41		35	36	37	38	39	40	41
Age at Immigration	0	139	135	162	151	183	172	172		758	679	838	834	954	858	910
	1	275	249	282	282	375	410	358		1,418	1,244	1,422	1,427	1,861	[illegible]	1,773
	2	368	309	333	329	338	483	488		1,886	1,531	1,809	1,659	1,758	[illegible]	[illegible]
	3	424	360	289	315	364	335	456		[illegible]	1,861	1,502	1,574	1,883	1,700	[illegible]
	4	507	364	350	309	346	331	368		[illegible]	1,868	1,759	1,557	1,809	1,700	1,854
	5	427	493	411	383	289	327	368		[illegible]	[illegible]	[illegible]	[illegible]	1,493	1,699	1,888
	6	294	435	459	388	345	304	321		1,509	[illegible]	[illegible]	1,961	1,774	1,553	1,679
	7	259	292	388	431	415	299	296		1,335	1,464	1,942	[illegible]	[illegible]	1,543	1,565
	8	221	226	316	416	513	386	341		1,156	1,195	1,574	[illegible]	[illegible]	1,977	1,707
	9	250	231	198	290	406	419	398		1,323	1,213	986	1,520	[illegible]	[illegible]	1,924
	10	366	293	238	246	324	393	446		1,922	1,505	1,221	1,288	1,615	1,986	[illegible]
	11	245	271	272	212	240	301	384		1,257	1,316	1,492	1,126	1,220	1,632	[illegible]
	12	252	247	362	304	240	206	308		1,276	1,258	1,968	1,554	1,283	1,059	1,502
	13	200	240	245	314	299	232	208		1,074	1,217	1,327	1,546	1,557	1,150	1,096
	14	228	228	224	250	356	302	221		1,165	1,153	1,226	1,338	1,834	1,496	1,125
	15	229	204	213	231	255	353	365		1,188	1,071	1,111	1,191	1,352	1,794	1,872
	16	252	217	215	212	285	305	392		1,284	1,093	1,059	1,049	1,454	1,628	[illegible]
	17	348	304	254	254	238	317	319		1,774	1,572	1,274	1,385	1,160	1,590	1,669
	18	393	415	340	330	286	282	337		[illegible]	[illegible]	1,818	1,734	1,528	1,556	1,777
	19	491	421	445	379	346	315	338		[illegible]	[illegible]	[illegible]	1,889	1,790	1,662	1,793
	20	554	579	559	468	414	410	426		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

		Sample Count								Population Total						
Age		42	43	44	45	46	47	48		42	43	44	45	46	47	48
Age at Immigration	0	144	127	102	118	128	142	178		726	723	517	590	611	698	956
	1	280	186	155	148	215	209	229		1,462	971	832	724	1,012	1,124	1,151
	2	403	349	283	184	157	201	224		[illegible]	1,829	1,383	939	775	1,054	1,090
	3	458	361	293	210	214	169	180		[illegible]	1,819	1,454	1,105	1,116	849	923
	4	483	470	365	310	215	158	150		[illegible]	[illegible]	1,919	1,497	1,124	812	791
	5	368	442	430	378	289	193	157		1,769	[illegible]	[illegible]	1,886	1,434	990	767
	6	348	315	404	460	340	232	193		1,719	1,653	[illegible]	[illegible]	1,666	1,224	1,004
	7	249	339	317	433	367	282	280		1,273	1,760	1,655	[illegible]	1,853	1,491	1,385
	8	313	259	288	329	381	353	315		1,586	1,269	1,493	1,642	1,914	1,810	1,676
	9	285	243	276	262	258	356	352		1,416	1,223	1,431	1,309	1,317	1,735	1,760
	10	391	361	287	287	274	273	320		[illegible]	1,837	1,469	1,399	1,314	1,362	1,693
	11	416	301	265	224	217	268	234		[illegible]	1,602	1,337	1,160	1,102	1,348	1,196
	12	399	407	338	283	193	243	254		[illegible]	1,975	1,724	1,461	1,004	1,222	1,252
	13	304	401	361	316	261	183	207		1,561	[illegible]	1,861	1,604	1,321	891	1,054
	14	256	275	381	368	326	265	204		1,318	1,439	1,952	1,822	1,674	1,383	1,049

	15	247	180	260	400	375	305	252	1,261	**923**	1,317	1,959	1,815	1,554	1,294
	16	420	280	244	265	338	392	290		1,528	1,265	1,356	1,761		1,437
	17	407	399	288	262	281	380	457		1,933	1,550	1,321	1,487	1,931	
	18	377	533	485	302	315	343	401	1,924			1,604	1,589	1,769	
	19	390	420	552	567	382	285	372	1,997				1,978	1,473	1,954
	20	404	441	458	681	654	443	397							

		Sample Count							Population Total						
	Age	49	50	51	52	53	54	55	49	50	51	52	53	54	55
Age at Immigration	0	221	143	144	150	128	147	122	1,180	**688**	**763**	**719**	**610**	**706**	**621**
	1	362	381	223	229	233	307	353	1,811	1,922	1,106	1,197	1,145	1,507	1,749
	2	286	453	492	259	299	388	421	1,386			1,263	1,627	1,917	
	3	244	279	428	459	274	341	378	1,256	1,413			1,287	1,705	1,905
	4	223	214	262	429	442	304	324	1,139	1,071	1,289			1,547	1,581
	5	158	206	219	249	382	450	322	**767**	1,039	1,073	1,279	1,964		1,640
	6	134	130	206	198	221	384	458	**639**	**659**	1,010	1,086	1,101	1,883	
	7	201	169	138	192	238	227	427	1,019	**851**	**682**	**964**	1,210	1,135	
	8	278	198	136	122	197	197	220	1,347	**963**	**742**	**656**	1,051	**950**	1,092
	9	284	246	183	160	110	162	192	1,441	1,252	**913**	**813**	**577**	**820**	**947**
	10	369	286	230	201	157	129	240	1,889	1,465	1,260	**979**	**761**	**672**	1,201
	11	302	330	274	215	158	137	88	1,556	1,657	1,345	1,090	**838**	**712**	**412**
	12	268	338	300	259	224	165	123	1,366	1,618	1,477	1,278	1,127	**883**	**608**
	13	221	223	324	280	236	223	185	1,107	1,128	1,635	1,465	1,095	1,180	**988**
	14	218	227	246	274	269	238	209	1,102	1,200	1,224	1,385	1,368	1,203	1,087
	15	177	227	242	241	293	281	269	**914**	1,209	1,238	1,234	1,491	1,501	1,393
	16	281	217	242	280	270	292	304	1,454	1,102	1,228	1,390	1,370	1,484	1,582
	17	397	326	298	303	308	297	353		1,680	1,546	1,525	1,624	1,489	1,801
	18	461	447	422	372	390	415	432				1,925	1,997		
	19	498	556	551	512	418	465	471							
	20	481	645	717	611	795	501	569							

Note: Since generally, the size of the population appears to be between 1,000 and 2,000 for most cells, to assist in determining possible periods of larger or smaller influxes, cells with population size lower than 1,000 are highlighted in bold, and cells with a population size of higher than 2,000 are highlighted by a grey background.
Source: 2006 Canadian Census of Population.

9.2 Language Score Table from Chiswick and Miller (2005)

Table 10: Index of difficulty of learning a foreign language (language scores) and codes for languages reported in the US Census

Language	Direct codes 1990, 2000 Censuses	Close codes 1990 Census	Changes for 2000 Census	Language score
Afrikaans	611			3.00
Danish	615			2.25
Dutch	610	612		2.75
French	620	621, 622, 623, 624		2.50
German	607	608, 609, 613		2.25
Italian	619			2.50
Norwegian	616	617, 618		3.00
Portuguese	629	630		2.50
Rumanian	631	632		3.00
Spanish	625	626, 627		2.25
Swedish	614			3.00
Indonesian	732	730-731, 733-737		2.00
Malay	739			2.75
Swahili	791	792		2.75
Amharic	780			2.00
Bengali	664			1.75
Bulgarian	647	648		2.00
Burmese	717			1.75
Czech	642			2.00
Dari	660			2.00
Farsi	656	657, 658, 659, 661		2.00
Finnish	679	680		2.00
Greek	637			1.75
Hebrew	778			2.00
Hindi	663	662, 665-669, 678	Add 671	1.75
Hungarian	682			2.00
Lao	720			1.50
Cambodian	726			2.00
Mongolian	694	695, 716		2.00
Nepali	674			1.75
Polish	645	644, 646		2.00
Russian	639	640, 641		2.25
Serbo-Croatian	649-651	652		2.00
Sinhala	677			1.75
Tagalog	742	740, 741, 743-749		2.00
Thai	720	717, 718, 719	Add 725	2.00
Turkish	691	689, 690, 692, 693		2.00
Vietnamese	728	729		1.50
Arabic	777	779		1.50
Mandarin	712	713, 714, 715		1.50
Japanese	723	725	Delete 725	1.00

Language	Direct codes 1990, 2000 Censuses	Close codes 1990 Census	Changes for 2000 Census	Language score
Korean	724			1.00
Cantonese	708	709, 710, 711, 721, 722		1.25

Note: Language codes in this table are from the 1990 US Census of Population and Housing, Technical Documentation and from the 2000 US Census of Population and Housing, Technical Documentation. There are minor differences in the language codes in the 1990 and 2000 Censuses. These differences are indicated in column (3). Column (4) is the language score for the direct codes.

Source of matching codes: (a) Grimes and Grimes (1993), (b) Adam Makkai, Professor of Linguistics, Department of English, University of Illinois at Chicago.

Source of Language Score: Hart-Gonzalez and Lindemann (1993).

9.3 Language Categories Definitions

Table 11: Language Category (English)

English	
Code	**Country**
17	USA
26	Anguilla
27	Antigua and Barbuda
29	Bahamas
30	Barbados
31	Bermuda
32	Cayman Islands
34	Dominica
36	Grenada
39	Jamaica
41	Montserrat
44	Saint Kitts and Nevis
45	Saint Lucia
46	Saint Vincent and the Grenadines
47	Trinidad and Tobago
49	British Virgin Islands
59	Guyana
89	Ireland (Eire)
95	United Kingdom
167	Republic of South Africa
221	Australia
231	New Zealand

Table 12: Language Category (French)

French	
Code	**Country**
16	Saint Pierre and Miquelon
37	Guadeloupe
38	Haiti
40	Martinique
67	France
140	Réunion
162	Gabon

Table 13: Language Category (Low)

Low	
Code	**Country**
99	Greece
179	Palestine/West Bank/Gaza Strip
194	People's Republic of China
195	Hong Kong *
196	Macau *
197	Japan
198	North Korea
199	South Korea
201	Taiwan
202	Brunei Darussalam
206	Laos
212	Viet Nam
213	Bangladesh
215	India
223	Fiji

* Special Administrative Unit

Table 14: Language Category (Medium)

Medium	
Code	**Country**
74	Bulgaria
75	Czech Republic
76	Slovakia
78	Hungary
79	Poland
81	Estonia
91	Finland
106	Bosnia and Herzegovina
107	Croatia
108	Macedonia
109	Serbia and Montenegro
110	Slovenia
111	Yugoslavia, n.o.s.
132	Eritrea
169	Afghanistan
171	Iran
185	Turkey
209	Philippines

n.o.s. not otherwise specified

Table 15: Language Category (High)

High	
Code	**Country**
19	Costa Rica
20	El Salvador
21	Guatemala
22	Honduras
23	Mexico
24	Nicaragua
25	Panama
33	Cuba
35	Dominican Republic
43	Puerto Rico
51	Argentina
52	Bolivia
53	Brazil
54	Chile
55	Colombia
56	Ecuador
60	Paraguay
61	Peru
63	Uruguay
64	Venezuela
65	Austria
68	Germany
69	Liechtenstein
72	Netherlands
80	Romania
84	Belarus
85	Republic of Moldova
86	Russian Federation
87	Ukraine
88	USSR, n.o.s.
90	Denmark
92	Iceland
93	Norway
101	Italy
103	Portugal
105	Spain
155	Angola
186	Kazakhstan
187	Kyrgyzstan
188	Tajikistan
189	Turkmenistan
190	Uzbekistan

n.o.s. not otherwise specified

9.4 Adjustment of Employment Income – Regression Output

Table 16: Regression results - Adjustment of Employment Income

Variable	Parameter	Standard Error	T-stat	P-value	Significance
Intercept	36,744	594	61.83	0.0000	**
NL	10,992	2,652	4.14	0.0000	**
PE	-8,004	3,283	-2.44	0.0148	*
NS	1,243	1,217	1.02	0.3070	
NB	-3,738	1,428	-2.62	0.0088	**
QC	-6,258	620	-10.09	0.0000	**
ON	3,432	555	6.19	0.0000	**
MB	-3,508	827	-4.24	0.0000	**
SK	-25	1,313	-0.02	0.9849	
AB	5,994	640	9.37	0.0000	**
BC	-128	588	-0.22	0.8272	
incma	6,293	366	17.18	0.0000	**
agedev	159	20	8.02	0.0000	**
agedev2	-46	4	-12.57	0.0000	**

Notes:

* Significant at the 5% level

** Significant at the 1% level