Microdata User Guide

LONGITUDINAL SURVEY OF IMMIGRANTS TO CANADA

WAVE 2

Statistics Canada Statistique Canada

Canada

Table of Contents

1.0 Introduction

The Longitudinal Survey of Immigrants to Canada (LSIC), conducted jointly by Statistics Canada and Citizenship and Immigration Canada under the Policy Research Initiative, is a comprehensive survey designed to study the process by which new immigrants adapt to Canadian society.

This guide was developed to facilitate use of the microdata file containing the results of Wave 2 of the Longitudinal Survey of Immigrants to Canada (LSIC). This file contains data from the first two collection waves. The collection was conducted by Statistics Canada, with the first wave taking place between April 2001 and May 2002 and the second between December 2002 and December 2003.

Any question about the data set or its use should be directed to:

Statistics Canada

Client Services
Special Surveys Division
Telephone: (613) 951-3321 or call toll-free 1 800 461-9050
Fax: (613) 951-4527
E-mail: ssd@statcan.ca

2.0 Background

The Longitudinal Survey of Immigrants to Canada is a comprehensive survey designed to study the process by which new immigrants adapt to or integrate into Canadian society. As part of adapting to life in Canada, many immigrants face challenges such as finding suitable accommodation, learning or becoming more fluent in one or both of Canada's official languages, participating in the labour market or accessing education and training opportunities. The results of this survey will provide indicators of how immigrants are meeting these challenges and what resources are most helpful to their settlement in Canada. The survey also examines how the socio-economic characteristics of immigrants influence the process by which they integrate into Canadian society.

The topics covered by the survey include language proficiency, housing, education, foreign credentials recognition, employment, health, values and attitudes, the development and use of social networks, income, and impressions about life in Canada. The questions address respondents' situation before coming to Canada and since their arrival.

With the exception of the module on income - in which the person most knowledgeable about the subject is asked to respond - no interview may be conducted by proxy. Some modules also contain questions about members of the household, such as questions on employment, income or demographic characteristics, and on children, such as education questions. The unit of analysis for the survey is always the selected immigrant, referred to as the longitudinal respondent (LR), even though some questions are about the experience of other household members like the spouse/partner of the children.

3.0 Objectives

There exists a growing need for information on recent immigrants to Canada. While full integration may take several generations to achieve, the Longitudinal Survey of Immigrants to Canada (LSIC) is designed to examine the process during the critical first four years of settlement, a time when newcomers establish economic, social and cultural ties to Canadian society. To this end, the objectives of the survey are two-fold:

- to study how new immigrants adjust to life in Canada over time; and,
- to provide information on the factors that can facilitate or hinder this adjustment.

3.1 Advantages of a Longitudinal Survey

As a longitudinal survey, the LSIC presents certain advantages. Most importantly, by interviewing the same cohort of sampled immigrants on successive occasions, we are able to directly and more efficiently examine the settlement process than if we were to draw a different sample of immigrants from the same population at each interview. This gain in efficiency comes at a price, in that, in some instances, the assumptions that underlie conventional analytic models no longer hold, particularly when it is the time dependent response — for example, a jobless spell — and not the change itself that is of interest. In such instances, a single immigrant may contribute more than one observation to the analysis (repeated measures). Furthermore, due to the complexity of the LSIC design and weighting, other factors must be considered when analysing the data.

The second wave of interviews represents the first cycle of longitudinal data for the LSIC cohort[1]. The wealth of information collected — in particular the complete histories provided in entities consisting of event lists such as List of studies (ST), List of jobs (JB) and List of places where the LR lived (WL) — will allow researchers to examine the changes that have occurred in the lives of LSIC immigrants over their first two years in Canada, and study the impact these changes have had on their settlement process. For example, the recognition of credentials, the acquisition of more education and work experience might be used to examine labour market success.

A number of different types of analyses are possible using the longitudinal data. For example, a simple descriptive analysis might estimate the number of immigrants whose highest level of education changed between survey occasions. The individual change in highest level of education might in turn be used with other socio-demographic and economic data to model the probability that, in their first two years in Canada, the immigrant was able to find a job (logistic or probit regression); or, taking advantage of the employment histories provided in the "Employment" roster, this same information could be used to examine the length of time, say in weeks, that it took to find a job (sometimes called survival data, duration or event history analysis).

There is a great deal of literature dealing with the analysis of longitudinal data. Diggle et al (2002) examine some of the issues related to the analysis of longitudinal data. Korn and Graubard (1999) discuss the analysis of survey data, providing examples and considerations when analysing longitudinal data from surveys of complex design. Allison (1999) provides a practical guide to fitting survival data models using the SAS system; the extension of survival models to the case of complex survey data is examined in Lawless and Boudreau (2002). These books and papers by no means represent an exhaustive list, but provide the reader with some good initial references.

[1] Because information was collected retrospectively on all jobs, addresses and educational courses and programs, some longitudinal analysis was possible using the data collected at the first wave. See Renaud and Goldmann (2005) for example.

3.2 Longitudinal Analysis Using the Wave 2 Data Files

The Wave 2 data files have been structured to facilitate longitudinal analysis. The user need not merge information across waves: the complete Wave 1 and Wave 2 response profile has been provided for each immigrant who responded to the Wave 2 interview. Furthermore, each variable has been given a wave-specific name, making it easy to identify questions and content common to both waves. More information on the database structure is presented in Chapter 5.0; the naming convention of variables is provided in Section 5.4. Additionally, Chapter 6.0 addresses the longitudinal comparability of concepts measured in Waves 1 and 2.

For each respondent, there is single longitudinal weight variable, WT2L (found on the LR entity), which should be used in all analysis conducted on the Wave 2 longitudinal data. This weight can be thought of as the number of immigrants in the Wave 2 population of interest represented by the responding immigrant. The population of interest is those immigrants in the LSIC cohort who still resided in Canada at the time of the Wave 2 interview. The derivation of the weights is presented in Chapter 12.0; the use of the weights is discussed in Chapter 14.0.

Due to the complexity of the sample design and weight adjustments, the standard variance formulae used in some analysis software are not appropriate. For these reasons, special methods and tools are recommended when analyzing LSIC data; these are discussed in Chapter 15.0.

References

Allison, P.D. (1999). *Survival analysis using the SAS system: a practical guide*, SAS Institute, Cary, N.C.

Diggle, P., Heagerty, P., Liang, K., Zeger, S. (2002). *Analysis of Longitudinal Data*, Oxford University Press, New York.

Korn and Graubard (1999). *Analysis of Health Surveys*, Wiley, New York.

Lawless, J.F., Boudreau, C. (2002). Modelling and Analysis of Duration Data from Longitudinal Surveys. *Proceedings of Statistics Canada Symposium 2002*, Statistics Canada, 11-522-XCB.

Renaud, J., Goldmann, G. (2005). Événements internationaux et biographie. Les répercussions du 11 septembre 2001 sur l'établissement économique des immigrants au Canada et au Québec, *Recherches Sociographiques*, XLVI, 2: 281-299.

4.0 Concepts and Definitions

There are many variables and concepts that are critical to the analysis of the Longitudinal Survey of Immigrants to Canada (LSIC) data. The following is an explanation of the key concepts in the LSIC.

Census family: Refers to a married couple (with or without children of either or both spouses), a couple living common-law (with or without children of either or both partners) or a lone parent of any marital status, with at least one child living in the same dwelling. A couple living common-law may be of opposite or same sex. "Children" in a census family include grandchildren living with their grandparent(s) but with no parents present. A census family is also referred to as an "immediate family" in the survey.

Citizenship: The status of being a citizen, either native-born or naturalized, sharing equally in the rights, privileges and responsibilities belonging to each individual.

Common-law partner: The person who, though not legally married to the respondent, is living with the respondent as his/her spouse. This partner may be of the same or opposite sex.

Credentials: The highest level of education as above a high school diploma, professional or technical credentials and any other degrees, diplomas or certificates from outside Canada constitute education credentials.

> ***Fully Accepted:*** The employer/institution recognizes a credential as being legitimate within determined standards.
>
> ***Partially Accepted:*** The employer/institution partially recognizes a credential as being legitimate within determined standards.
>
> ***Not Accepted:*** Credential is not recognized as being legitimate within determined standards.

Discrimination: The unfavourable treatment of individuals on the basis of their personal characteristics, which may include race or skin colour, ethnicity or culture, language or accent, religion etc.

Economic family: Refers to a group of two or more persons who live in the same dwelling and are related to each other by blood, marriage, common-law or adoption.

Ethnic or Cultural Group: A group of individuals having a distinct culture in common. The term "ethnic or cultural group" implies that values, norms, behaviour and language, *not necessarily physical appearance*, are the important distinguishing characteristics.

FOSS: The acronym stands for "Field Operations Support System" and is an administrative database maintained by Citizenship and Immigration Canada. The FOSS was used as the sample frame for the survey.

Full-time Employment: Employment where people usually work 30 hours or more per week at their main or only job.

Host Program: This program matches newcomers with a volunteer who is familiar with Canadian ways, i.e. someone who can teach newcomers about available services, make contacts, help with employment, housing, etc. This program is intended to facilitate the integration process of newcomers.

Immigration categories:

***Economic Class*:** Includes immigrants selected for their skills or other assets that will contribute to the Canadian economy (includes skilled workers, investors, entrepreneurs, and self-employed persons).

Family class: Includes immigrants sponsored by close relatives or family members already living in Canada.

Independent immigrants: Includes immigrants who qualify for certain types of jobs or have other important assets to bring to Canada. They apply on their own or have more distant relatives living in Canada.

Refugees: Persons seeking protection in Canada**.**

Immigrant Settlement and Adaptation Program (ISAP): A program in which funds are provided to deliver direct and essential services to newcomers. These services include reception and orientation, translation and interpretation, referral to community resources, para-professional counselling, general information and employment-related services.

Immigration Consultant: A professional who gives advice or services related to immigration issues.

Immigrant or Refugee Serving Agency: An organized body catering to the needs of immigrants or refugees.

Immigrating Unit: Refers to a group of people who applied to come to Canada under the same visa form and, for the purpose of the survey, who arrived either with the longitudinal respondent or three months before or after the longitudinal respondent.

Immigration Officer: A Canadian official who processes the authorization of immigrants upon arrival in Canada.

Integration: The process through which newcomers participate in and shape Canadian community.

Joiner:

Wave 1: A person who was not a member of the longitudinal respondent's (LR) immigrating unit, but who was living in the same household at the time of the interview. This includes people who were already living in Canada when the LR arrived.

Wave 2: A person living in the longitudinal respondent's household but who was not a member of the longitudinal respondent's household at the previous wave. This includes people who were already living in Canada when the LR arrived

Longitudinal respondent (LR): The longitudinal respondent is the person selected to answer the LSIC questions at each of the three waves.

Mover: A person who was a member of the longitudinal respondent's immigrating unit, but who was not living in the same household at the time of the interview.

Part-time Employment: Part-time employment refers to persons who usually work less than 30 hours per week at their main or only job.

PMK: Person Most Knowledgeable about a specific subject. In the LSIC, the only questions asked of the PMK were questions on family income within the Income Module. If the PMK is not available, the questions are asked to the LR.

Population Group: Refers to the population group to which the respondent belongs. It includes visible minorities (see definition below) as well as Aboriginal peoples, Caucasian in race or white in colour.

Reference period: It is the time period in which a question in the survey fits. In other words, it is the date and the length of time to which a question is limited (i.e. the period of time covered by a question). E.g.: period of time between the first and the second interview. Reference periods may change from one wave to the next for the same question.

Sponsor: Canadian Citizens, or permanent residents aged 19 or over, living in Canada that commit to provide the sponsored immigrant with basic assistance in the form of accommodation, clothing, food and settlement assistance for a specific period of time.

Visible Minority: Refers to persons, other than Aboriginal peoples, who are non-Caucasian in race or non-white in colour.

5.0 File Structure and Content

The data file for the second wave of the Longitudinal Survey of Immigrants to Canada (LSIC) contains data from the first two collection waves of the survey. It contains all records relating to the 9,322 respondents who were traced and agreed to respond to both waves. Having the data from both waves merged together will facilitate longitudinal analysis of the Wave 1 and Wave 2 survey results.

5.1 Data Model

The LSIC data have been divided into a number of smaller databases, called entities. This structure, which is called the data model, represents an intuitive and practical way of storing longitudinal data. Each entity includes variables relating to the same concept that largely reflects the structure of the questionnaire modules. The following figures and tables show the LSIC entities and describe their content.

Figure 5.1 Entities that have one record per respondent

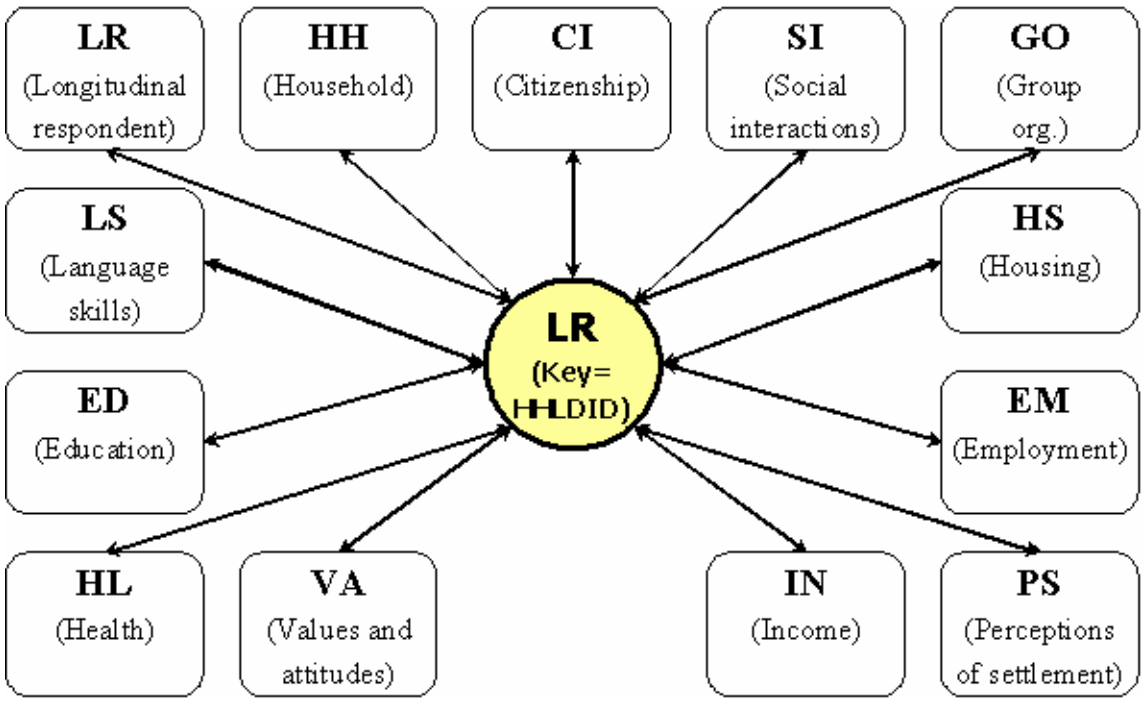


Figure 5.2 List of entities that can contain more than one record per respondent

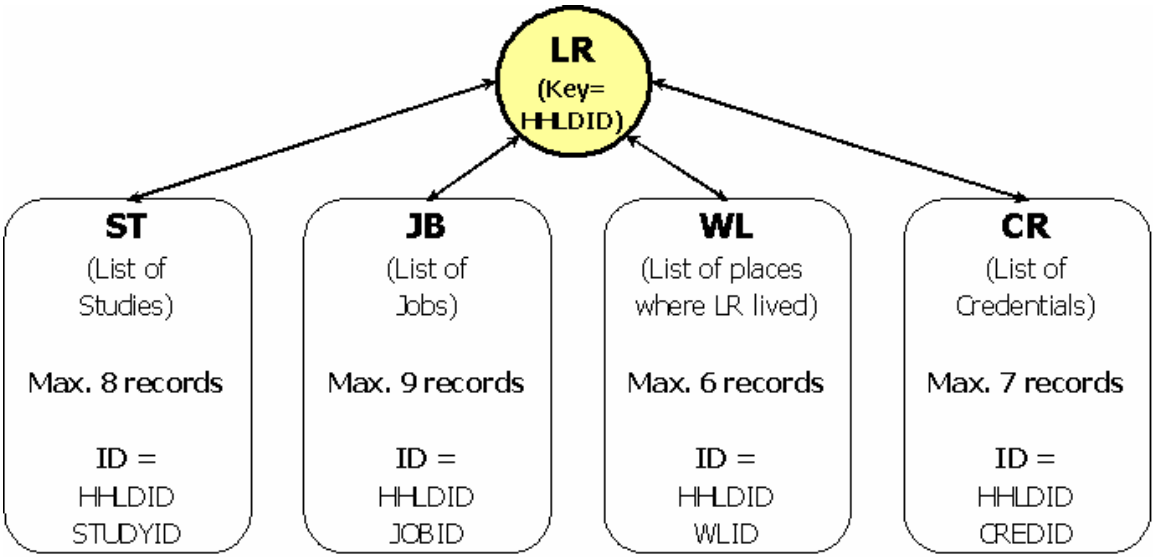


Table 5.1 List of the data model entities and their contents

Acronym	Entity Name (concept)	Unique Key Identifier	Wave 1 Collected or Derived From:	Wave 2 Collected or Derived From:
CI	Citizenship	HHLDID	Background Module: BG_Q06 to BG_Q09B, BG_Q16 and BG_Q17 and Values and Attitudes Module: VAS_Q01 to VAS_Q04A	Citizenship Module
CR	List of education credentials	HHLDID, EDCID	Education Credentials (sub-module of the Education Module)	Education Credentials Module (sub-module of the Education Module)
ED	Education	HHLDID	Education Module	Education Module
EM	Employment	HHLDID	Employment Module	Employment Module
GO	Groups and organizations	HHLDID	Group Organizations (sub-module of the Social Network Module)	Group Organizations (sub-module of the Social Network Module)
HH	Household	HHLDID	Entry Module (includes aggregated derived variables from the relationships questions)	Entry Module (including the household relationship matrix)
HL	Health	HHLDID	Health Module	Health Module
HS	Housing	HHLDID	Housing Module, Background Module: BG_Q14 and BG_Q15	Housing Module
IN	Income	HHLDID	Income Module	Income Module

Acronym	Entity Name (concept)	Unique Key Identifier	Wave 1 Collected or Derived From:	Wave 2 Collected or Derived From:
JB	List of jobs	HHLDID, JOBID	Employment Details and Employment Roster (sub-modules of the Employment Module)	Employment Details and Employment Roster (sub-modules of the Employment Module)
LR	Longitudinal respondent	HHLDID	Entry Module, Background Module: BG_Q01 to BG_Q05 and BG_Q18 to BG_Q20, in addition to some variables from a Citizenship and Immigration Canada administrative database.	Entry Module
LS	Language skills	HHLDID	Language Skills Module excluding language test: LS_Q11E to LS_Q16E and LS_Q11F to LS_Q16F	Language Skills Module
PS	Perceptions of settlement	HHLDID	Perceptions of Settlement Module	Perceptions of Settlement Module
SI	Social interactions	HHLDID	Social Network Module	Social Interactions Module
ST	List of studies	HHLDID, STUDYID	Education Details and Education Roster (sub-modules of the Education Module)	Previous Education, Education Roster and Education Details (sub-modules of the Education Module)
VA	Values and attitudes	HHLDID	Values and Attitudes Module, excluding VAS_Q01 to VAS_Q04A	Values and Attitudes Module
WL	List of places where the LR lived	HHLDID, WLID	Where Lived (sub-module of the Housing Module)	Where Lived (sub-module of the Housing Module)

5.2 File Format

The LSIC files are available in two formats:

1. **Text files (in ASCII format)**
   All entities are included in one large text file (MAIN) except for the entity containing information on the respondent's household and the roster entities. The Household entity (HH) and the roster entities (CR, JB, ST, WL) each have their own separate text file. SAS and SPSS syntax cards have been provided for formatting these files (names of these files end by SASE and SPSSE for English syntax cards and SASF and SPSSF for French syntax cards).

Table 5.2 Text File Structures

File Name	File Description
LSIC_W2_MAIN_Master.txt	This file includes the following entities: LR, CI, SI, GO, HS, HL, LS, ED, EM, VA, IN, PS.
LSIC_W2_HH_Master.txt	This file includes information collected on the respondent's household.
LSIC_W2_CR_Master.txt	This file includes variables collected in the Education Credentials sub-module.
LSIC_W2_JB_Master.txt	This file includes variables collected in the Employment Roster and Employment Details sub-modules.
LSIC_W2_ST_Master.txt	This file includes variables collected in the Education Roster and Education Details sub-modules.
LSIC_W2_WL_Master.txt	This file includes variables collected in the Where Lived sub-module.

2. **Files in SAS format**
   Each entity constitutes an individual file as described in Table 5.1. All LSIC files include a unique key identifier referred to as the household identifier (variable name HHLDID) which is specific to the longitudinal respondent (LR). All LSIC files can all be merged using this unique key variable. All roster entities also contain other identifiers to make each record unique.

5.3 File Content

All entities except roster entities contain one record per longitudinal respondent who provided responses to the first two waves of the LSIC, for a total of 9,322 records. Wave 2 variables have unique names, and these variables were added to the Wave 1 variables to form the Wave 2 files.

The roster entities, namely Education Credentials (CR), List of Jobs (JB), List of Studies (ST) and List of Places Lived (WL) may contain more than one record per respondent. In these entities, the minimum number of records for a respondent is zero and the maximum collected varies depending on the entity (CR = 7, JB = 9, ST = 8, WL = 6).

It is important to note that when producing estimates, the final weights are only to be used for the LR records. No weighted estimate can be produced directly from roster entity records. For further details, please refer to Chapter 12.0 on weighting.

As described above, files for the second wave of the LSIC contain data collected in the first two waves. In other words, each entity contains both the Wave 1 and the Wave 2 variables.

The entities contain a number of derived variables. Most derived variables are located at the end of files (entities).

5.4 Structure of Variables

To facilitate interpretation of the data by users, the assignment of variable names and values is governed by certain rules in the documentation system for the LSIC microdata file. First, each variable name contains an identifier (item identifier - see Chapter 6.0 on longitudinal

comparability) which serves to identify variables that are longitudinally linked from one wave to the other, that is, variables that measure similar phenomena but at different periods.

All variable names are, at most, eight characters long (most are 7 long) so that these names can easily be used with analytical software packages such as SAS or SPSS.

Description of the structure of variable names:

- The **first two** characters are the acronym of the entity to which the item belongs. See Table 5.1 for descriptions.

- The **third** digit of the variable name refers to the LSIC wave:

  "1" indicates the first wave,
  "2" indicates the second wave and
  "3" indicates the third wave.

- The **fourth** character provides information on the type of variable. There are six different types of variables.

  c Coded variable: A variable coded with standard exhaustive code sets (SOC91 - Standard Occupational Classification system, NAICS – North American Industry Classification System, CIP – Classification of Instructional Programs and the Census Country Code set).

  d Derived variable: A variable created from one or more collected or coded variables (e.g., household size, labour force status, etc.).

  g Grouped variable: Collected, coded or derived variables collapsed into groups (e.g., age groups, world regions, etc.).

  i Imputation flag: Indicates that the value of a variable for a respondent was imputed (field imputation), or that an entire entity was imputed (mass imputation). Field imputation flag variables directly follow the questions imputed. For example, the imputation flag variable for IN1Q003 is IN1I004.

  q Collected variable: A variable that contains the response to a question which was directly asked to the respondent.

  z Variables obtained from a linkage with administrative records from Citizenship and Immigration Canada.

- The **fifth, sixth and seventh** characters constitute a sequential number (starting at 001) assigned to each variable in the file. Within a given entity, this number remains the same from one wave to the next for any longitudinally linked variable.

  However, the order of the variables in the file does not correspond to this sequential number; instead, they tend follow a logical order based on the themes and the order of questions in the questionnaire. Changes made to the Wave 2 questionnaire substantially modified the order initially anticipated in Wave 1.

- The **eighth** and final character (a letter) is used to indicate important changes to a variable from one wave to another that could affect the comparability of the two variables. If a change has the effect of altering the meaning of a question or the values associated with it, the variable is treated as new and an "x" is attached. The decision as to whether a new variable needs to be created (renamed) is discussed in Chapter 6.0.

Table 5.3 Examples of variable names

Example 1: Variable CI1Q002	
CI	Variable from the Citizenship entity
1	Wave 1 variable
Q	Taken directly from a question (included in the questionnaire)
002	Variable number 002 from the Citizenship entity

Example 2: Variable HL2D004x	
HL	Variable from the Health entity
2	Wave 2 variable
D	Derived variable
004	Variable number 004 from the Health entity
x	Means that this variable is similar but not identical to the variable HL1Q004. In this case, the wording was changed in the Wave 2 questionnaire. The change was considered sufficiently important for a new variable to be created. Note that this change is identified in the concordance table (see Section 6.5 for more information on the Concordance table).

6.0 Longitudinal Comparability

6.1 Changes in the Questionnaire

In longitudinal surveys, a general rule is that the questions must remain identical from one wave to the next with the exception of the reference period. Thus, the variables created in each wave measure the same phenomenon, but at different times, which allows users to conduct longitudinal analyses.

However, the questionnaire for the second wave of the Longitudinal Survey of Immigrants to Canada (LSIC) underwent many revisions. Some questions were dropped and others were subjected to major changes. While these changes are generally intended to make the questions more understandable, they may affect longitudinal comparability.

6.2 Nomenclature: Items and Variables

For clarity purposes, a nomenclature was developed that is very practical for grasping the longitudinal aspect of the variables. A distinction is made between what are called “items” and “variables.” An *item* is a particular phenomenon, measured specifically, in a given set of respondents for a specific reference period. A *variable* is the representation or measurement of an *item* in a wave.

The identifier of the *item* is embedded in the name of the *variable*, so that they can be linked intuitively. The first two characters (which identify the entity), when combined with the fifth, sixth and seventh characters (which identify the *variable* within the entity), serve to identify the *item*, and they will always remain invariable from one wave to the next. For example, variable HH1Q009 measures the number of persons in the household as reported in Wave 1, while *variable* HH2Q009 measures the same thing as reported in Wave 2. We will therefore speak of *variables* HH1Q009 and HH2Q009, and more generally of *item* HH_009, which represents the number of persons in the household in a wave.

6.3 Scope of Changes

Changes made to the questionnaire may jeopardize longitudinal comparability. Minor changes slightly affect the way an *item* is measured compared to the previous wave. It may be a matter of words used in the questionnaire or additional directions given to the interviewer. To enable users to judge the impact of the changes to the questionnaire for their analysis, all changes, even minor, are indicated in the concordance table (see Section 6.5 on the concordance table).

However, some major changes in the Wave 2 questionnaire are such that the *variables* that result from them can no longer be associated with *items* that already existed in Wave 1. Some *items* then become obsolete, and it was necessary to create new ones. This was done so that users can recognize longitudinal comparisons that may be suspect, namely comparisons between *variables* measuring different *items*.

When a new *item* similar to an existing one must be created, it retains a similar identifier. The new *item* takes on the same identifier as the similar *item* that it replaces, except that an ”x” is added at the end, becoming the eighth character in the name of the variable. In this context, an “x” at the end of a variable may be seen as an indicator that a major change occurred between Wave 1 and Wave 2. For example, variable EM1Q049 is based on question EM_Q19 of the Wave 1 questionnaire, and it measures the main activity of the respondent in Wave 1. In Wave 2, the main activity is measured by question EM_Q02. However, question EM_Q02 of Wave 2 differs substantially from question EM_Q19 of Wave 1 since new response categories have been added. For this reason, these two variables are said to measure two *items* that are different but similar: EM_049 in Wave 1 and EM_049x in Wave 2.

This convention serves to preserve the intuitive link between two *items* that are similar but not identical. An informed user who has to use a variable containing an "x" in the eighth character will seek to understand the nature of the changes made. The concordance table should be the first stop, informing the user about the differences between the *items* (see following section).

Entities consisting of event lists – namely, List of places where the longitudinal respondent (LR) lived (WL), List of studies (ST), List of education credentials (CR) and List of jobs (JB) – are exceptions, since the eighth character of the *variable* is often used to identify events when a flat file (one record per respondent) is created. For these entities, then, new *variables* received a new *item* identifier. However, in these cases, a note in the concordance table indicates the link with a *variable* from the previous wave.

6.4 Nature of Changes

Conceptually speaking, six types of changes in the questions or in the ways of asking questions from one wave to the next may affect longitudinal comparability:

1) Wording or concept
Sometimes, changes in the wording of questions are intended to measure a phenomenon that is similar but not identical to what was measured in the previous wave. Other times, the purpose is merely to clarify questions that may have created problems for respondents during collection. However, nuances in the wording of questions may introduce differences in the understanding of the phenomenon and in response behaviour.

2) Instructions to interviewers
Some questions involve specific instructions for interviewers. A change in the instructions can lead to different response behaviour. In most cases, these are questions where the interviewer must read response choices to respondents for a question in one wave but not in another.

3) Response categories
This type of change affects only those questions where respondents can provide only a single response from among a choice of several responses. In general, the response categories remain identical in the questionnaires from one wave to another. However, some questions have undergone changes. Most of the time, the change consists of adding categories so as to obtain more details. When it was not possible to recreate an item by manipulating response categories, a new item was created.

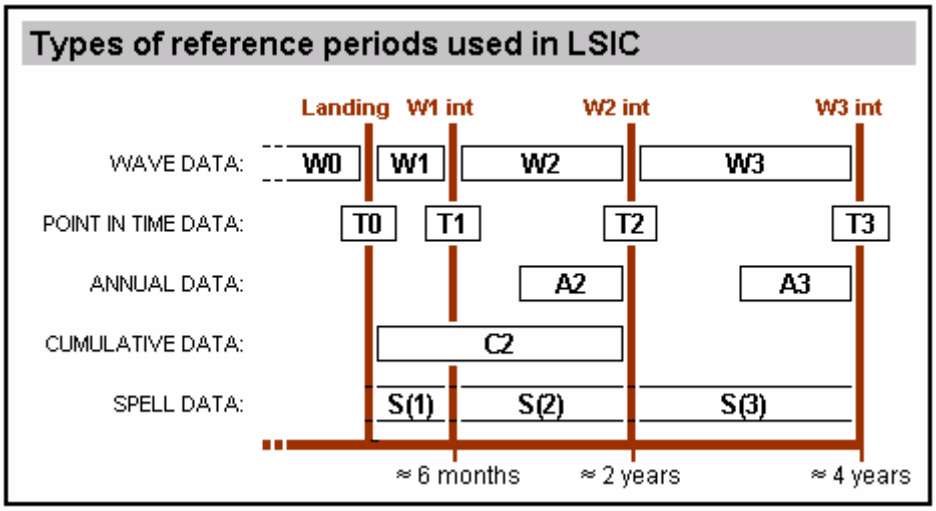
4) Universe
The universe, or coverage, consists of those immigrants to whom the data of a variable apply. It is therefore sensitive to sequencing in the questionnaire. Most of the time, a change in the universe also causes a new item to be created.

5) Structure of questions
Sometimes, information collected in Wave 1 may be collected differently in Wave 2 without there being a change of concept or a change in the universe. Often this involves a change in the structure of questions.

An example of this is when, to ensure longitudinal comparability, two questions in the second wave are combined to create a derived variable measuring an existing item in Wave 1.

6) Type of reference period
It is important to understand the temporal dimension that is central to longitudinal surveys. The LSIC variables can be categorized into five types according to their relationship to time:

wave variables, annual variables, variables referring to a point in time, event variables and cumulative variables.



Two variables measuring an item in different waves will have different reference periods. On the other hand, the type of reference period for the two variables will be the same. In other words, for a given item, the reference period changes for each variable from one wave to the next, but the type of reference period remains unchanged. If, because of changes in the questionnaire, it is no longer possible to measure an item for a given type of reference period, a new item is created. Below is a description of the different types of reference periods, along with an example of a question for each:

Wave variables
These entirely cover the period between the date of arrival in Canada and the Wave 1 interview, or between two consecutive interviews. In this case, the length of the reference period is unequal from one wave to another, ranging from approximately six months in Wave 1 to a year and a half in Wave 2. Example: *Since your last interview, have you received any medical attention?*

Annual variables
Annual variables cover a 12-month period preceding the date of the interview. It should be noted that there were no variables of this type in the first wave. Example: *In the last 12 months, did you receive income from sources within or outside Canada?*

Variables relating to a point in time
These reveal a situation that exists at the time of the interview. Example: *How many rooms are there where you live?*

Variables relating to an event
These refer to events that took place between a starting date and an ending date. The duration of these events varies; it may extend beyond the collection waves and is not pre-specified in the questionnaire. This type of variable is mainly found in entities consisting of event lists such as List of studies (ST), List of jobs (JB) and List of places where the LR lived (WL). Example: *Why did you not complete this course or program?*

Cumulative variables
These cover both waves. They are fairly rare in the LSIC. An example is this question asked in Wave 2: *Since your arrival in Canada, how often have you experienced discrimination or unfair treatment?*

In short, changes to the type of reference period are rather rare, except in the case of income variables. In that case, respondents were asked in the first wave to provide the income earned since coming to Canada. Thus the income variables here were wave variables. In the second wave, respondents are asked for income earned in the 12 months preceding the interview date. The income variables here are therefore annual variables.

6.5 Making Connections: the Concordance Table

The concordance table indicates not only whether there has been a change but also the nature of the change(s). Each line represents a specific item. It is thus easy to see which items have been dropped and which ones have been added. Each item has a short description that uniquely defines it. This description is identical to labels associated to each variables contained in the syntax cards (predefined formats).

The "Note" column shows one or more codes indicating the nature of the changes, or any comments regarding the use of a variable. Users can consult the data dictionary to obtain further details on differences between waves. The following codes have been assigned to identify the nature of changes as described in Section 6.4.

Code	Change
Wm	Wording/meaning
In	Instructions to interviewers
Rc	Response categories
Un	Universe
St	Structure of questionnaire
Pt	Reference period

Note that the concordance table indicates changes in the way an item is measured in a wave but also changes that led to the creation of a new item.

6.6 Index of Major Changes

The following is a list of the most important changes to the questionnaire in Wave 2 of the LSIC:

1) Type of reference period for income variables
In the first wave, respondents had to report amounts received from different sources since coming to Canada. In the second wave, respondents must report amounts received for the twelve months preceding the interview date. Collecting income on an annual basis will allow comparisons with many other data sources, since most surveys collect annual income.

2) Separation of language courses and other types of education
In the first wave, the education module collected data on all types of courses taken by the respondent. To obtain specific details on language courses, it was decided that language courses should be separated from other types of training in the second wave.

The consequences are two-fold. First, some questions in the Education module were repeated in the Language Skills (LS) module in order to obtain information on problems obtaining language training. Second, the Education Roster and Details (ST) module that was used to collect information on all courses taken by the respondent in the first wave was also changed. In the second wave, this module collects information on all courses taken by the respondent except language courses. On the other hand, some questions were added to the Language Skills (LS) module in order to obtain details on language courses. While it is possible to determine whether the respondent took language courses to learn or improve

his/her English or French and also to find out various details concerning these courses, it is no longer possible to determine the number of courses taken and the starting and ending dates of the courses (and thereby the duration).

3) Random selection of a child in the household

Three modules -- Education, Health, and Values and Attitudes -- include questions about the LR's children. For example, questions in the Health module look at whether children have experienced dental problems while in Canada (see questions HL_Q08B and HL_Q29 in Waves 1 and 2 respectively). In Wave 1, respondents were asked questions about their children in general. In Wave 2 (and again in Wave 3), to obtain more specific information, one child was selected at random from all of the LR's children aged 2 to 18 years. While the questions refer to the selected child in Waves 2 and 3, and all children in Wave 1, the unit of analysis remains the LR, i.e. child data are to be used as attributes of the LR in all waves. However as a result of the change in methodology, it might not be possible to make general inference about the Wave 2 and 3 populations of interest based on some of the selected child questions. For example, using question HL_Q29 to estimate the number of immigrants who were parents of a child who had dental problems will result in an underestimate, as a respondent for whom the selected child had no such problems might have another child who did. The user is thus cautioned as to the interpretation of analyses involving the child questions.

4) The addition of filter questions

Several Wave 1 questions attempted to force an answer from respondents. For example, respondents were asked: *What problems or difficulties have you had finding a job in Canada?* without having already been asked whether they had problems finding a job.

In Wave 2, questions of this type were reworded. For each of those, a filter question was added. In the above example, respondents were asked: *Since your last interview, have you had any problems or difficulties in finding a job in Canada?*. Those who answered that they had had problems were asked what problems.

7.0 Sample Selection

The Longitudinal Survey of Immigrants to Canada (LSIC) was designed to collect longitudinal data on immigrants in order to better understand the process by which new immigrants adapt to Canadian society. This survey will provide information on factors which facilitate or impede their adaptation and the ways that they contribute to Canadian society and the Canadian economy.

The completed survey will consist of three interviews (waves): the first (Wave 1) of these was conducted six months after the immigrant's arrival in Canada; the second (Wave 2), 2 years after arrival; with the third (Wave 3) occurring four years after their arrival.

To produce reliable estimates, a representative sample of approximately 20,300 new immigrants to Canada was selected. This chapter describes the selection of the LSIC sample.

7.1 Survey Populations

The ***target population*** for the survey consists of immigrants who meet all of the following criteria:

- arrived in Canada between October 1, 2000 and September 30, 2001;
- were age 15 or older at the time of landing;
- landed from abroad, must have applied through a Canadian Mission Abroad.

Individuals who applied and landed from *within* Canada are excluded from the survey. These people may have been in Canada for a considerable length of time before officially "landing" and would therefore likely demonstrate quite different integration characteristics to those recently arrived in Canada. Refugees claiming asylum from within Canada are also excluded from the scope of the survey.

The target population accounts for approximately 169,400[2] of the 250,000 persons admitted to Canada during this period. Coverage of the survey included all Census Metropolitan Areas and non-remote Census Agglomerations.

The ***population of interest*** is those immigrants in the target population who still reside in Canada at the time of a given wave. During the six months between arrival and the time of the first interview, and the period of time between the first and second interviews, some immigrants left Canada to return to their country of origin, or for another country, and are thus excluded from the population of interest. At Wave 1, this population was estimated at approximately 164,200 immigrants; at Wave 2, the size of the population of interest was estimated to be 160,800.

7.2 Survey Frame

The target population is represented by the survey frame from which the sample is selected. The sampling frame for the LSIC is an administrative database of all landed immigrants to Canada that comes from Citizenship and Immigration Canada. The database, known as the FOSS (Field Operation Support System), includes various characteristics of each immigrant that can be used for survey design purposes, such as: name, age, sex, mother tongue, country of origin, knowledge of English and/or French, class of immigrant, date of landing, and intended province of destination in Canada.

Detailed information from the FOSS on each immigrant landing during the survey reference period, i.e., October 2000 to September 2001, was provided to Statistics Canada two months

[2] Size of the target population according to an update to the survey frame; at the time of sample selection, approximately 165,000 immigrants were identified as belonging to the target population (see Table 7.1).

after the reference month. This allowed for the sampling frame to be built month after month by simply adding new monthly landings.

7.3 Survey Design

The survey was designed based on probability sample theory. The sample was created using a two-stage stratified sampling method. The first stage involved the selection of Immigrating Units (IU) using a probability proportional to size (PPS) method. The second stage involved the selection of one IU member within each selected IU. The selected member of the IU is called the longitudinal respondent (LR) and was contacted to participate in the survey. Only the LR is followed throughout the survey; no interviews are conducted with other members of the IU or the LR's household.

7.3.1 Longitudinal Sample

The survey involves a longitudinal design with immigrants being interviewed at three different times: at six months, two years, and four years after landing in Canada. The sample design has been developed using a "funnel-shaped" approach—i.e. a monotonic design—therefore only immigrants that responded to the Wave 1 interview were traced for the Wave 2 interview and only those that responded to the Wave 2 interview will be traced for the Wave 3 interview.

The funnel-shape approach was chosen because of the nature of the survey and its analytical objectives. The survey collects information on perceptions, values and attitudes at specific points in time, in order to assess the immigrant's integration during their initial years in Canada. If data were collected only once (i.e., during the fourth year in Canada), significant recall and response errors could be encountered. Furthermore, to facilitate a complete study of the immigrant's adaptation, the full range of longitudinal data must be obtained from each longitudinal respondent.

7.3.2 Stratification

The first stratification variable used was the month of landing in Canada; there are 12 cohorts of immigrants, i.e. one for each reference month. Within each month, two other stratification variables were used: the intended province of destination as stated by the immigrant and the class of immigrant.

Provinces were grouped into five categories: Québec, Ontario, Alberta, British Columbia and the remaining provinces (the territories were excluded).

For the purpose of stratification, immigrants were divided into six categories: family class, economic-skilled, economic-business, government-sponsored refugees, other refugees and other immigrants. Strata were created by the intersection of the above categories; thus, 30 strata were used for each monthly cohort of immigrants for a total of 360 strata.

7.4 Sample Selection and Sample Size

The sample was divided into two components - the core and the additional samples. The core sample represents the target population, while the additional samples target specific sub-populations. These specific sub-populations were determined by analysing the expected sample allocation at Wave 3 and also by various requirements of federal and provincial government departments. The following subgroups have been over-sampled:

1) government sponsored refugees;

2) refugees other than government sponsored;
3) contractor and investor immigrants (economic-business);
4) family immigrants in British Columbia;
5) overall immigrants in Alberta; and
6) economic immigrants in Québec (economic-skilled and economic-business).

The stratification allowed for control over the sample sizes for each of the additional samples' subgroups.

Tables 7.1, 7.2 and 7.3 provide a breakdown of the population based on the sampling frame and of the sample allocation for the core and additional samples expected at Wave 3.

For the core sample, it was determined that 5,000 completed interviews at Wave 3 would produce reliable estimates[3] at the national level; at the provincial level, where the in-flow of immigrants is the most significant (Québec, Ontario and British Columbia); and, for certain classes of immigrants (family and economic classes). Also it would be possible to obtain reliable estimates for other combinations of variables as long as a minimum number requirement is met. After taking into account the requirements for the additional samples outlined above, the minimum number of completed interviews at Wave 3 is expected to be 5,755 immigrants.

The determination of the sample size for Wave 1 was based on several sample attrition hypotheses applied to the Wave 3 minimum sample size requirement. Examining results from various longitudinal studies of the Canadian population, a combined response rate (resolved cases and respondent) of 75% was estimated for Waves 2 and 3 - i.e. 75% of Wave 1 respondents would respond in Wave 2 and 75% of Wave 2 respondents in Wave 3. In addition, various sources were used to estimate a combined return rate, i.e. after tracing and classification as in-scope or out-of-scope. Results from the pilot study and a coverage study on language[4] were used as a source of information. Finally, Statistics Canada's Reverse Record Check Study (RRC)[5] was used to estimate the expected tracing rates or rates of resolved cases.

The initial sample was selected over a 12-month period. A sample allocation proportional to the number of immigrants in each month of landing, as well as between strata within a month, would have minimized the total sampling variance. However, for operational reasons, such as maintaining a constant number of interviews in each month of collection, an equal allocation was performed between the months of landing, even though immigration shows a seasonal pattern. Table 7.4 presents the final sample size at Wave 1.

[3] By reliable estimates we mean being able to estimate a minimal proportion of 10% with a coefficient of variation of 16.5%. A cell size of 450 responding units is necessary to meet this requirement.

[4] Given operational constraints, namely the requirement and associated costs to translate the questionnaire in several languages, a study has been performed to identify the population coverage according to languages. It has been determined that the translation could be performed in 13 languages other than English or French, and that it would allow a national coverage of around 93% of landed immigrants.

[5] The 1996 RRC study was undertaken following the 1996 Census to estimate Census under-coverage. This study makes use of an immigrant frame that covers immigrants who landed in Canada between the 1991 and 1996 censuses.

Table 7.1 Total Number of Immigrants, 15 Years and Over, by Province and Class of Immigrant, October 2000 to September 2001

Province	Family	Economic-skilled	Economic-business	Government-refugee	Other Refugee	Other	Total
Quebec	4,680	12,694	2,977	1,238	887	78	22,554
Ontario	26,579	64,346	3,591	2,054	2,123	216	98,909
Alberta	3,250	5,651	444	623	307	125	10,400
British Columbia	8,532	15,048	2,489	679	317	235	27,300
Other provinces	1,199	2,074	494	948	427	707	5,849
Canada	**44,240**	**99,813**	**9,995**	**5,542**	**4,061**	**1,361**	**165,012**

Table 7.2 Expected Allocation of Respondents in Wave 3 - Core Sample

Province	Family	Economic-skilled	Economic-business	Government-refugee	Other Refugee	Other	Total
Quebec	151	312	94	46	25	5	633
Ontario	810	1,870	125	46	72	12	2,935
Alberta	104	156	21	13	6	4	304
British Columbia	287	505	108	12	10	10	932
Other provinces	41	74	19	25	12	25	196
Canada	**1,393**	**2,917**	**367**	**142**	**125**	**56**	**5,000**

Table 7.3 Expected Allocation of Respondents in Wave 3 - Core and Additional Samples

Province	Family	Economic-skilled	Economic-business	Government-refugee	Other Refugee	Other	Total
Quebec	151	346	125	146	28	5	801
Ontario	810	1,870	153	146	79	12	3,070
Alberta	154	231	36	47	9	6	483
British Columbia	450	505	132	38	11	10	1,146
Other provinces	41	74	23	79	13	25	255
Canada	**1,606**	**3,026**	**469**	**456**	**140**	**58**	**5,755**

Table 7.4 Final Sample Allocation at Wave 1

Province	Family	Economic-skilled	Economic-business	Government-refugee	Other Refugee	Other	Total
Quebec	463	1,230	437	377	111	12	2,630
Ontario	2,653	6,920	599	630	269	23	11,094
Alberta	531	928	93	234	59	22	1,867
British Columbia	1,560	1,634	423	210	40	26	3,893
Other provinces	121	225	81	293	46	72	838
Canada	**5,328**	**10,937**	**1,633**	**1,744**	**525**	**155**	**20,322**

8.0 Data Collection

8.1 Computer-assisted Interviewing

Data collection for the Longitudinal Survey of Immigrants to Canada (LSIC) relied heavily on computer-assisted interviewing (CAI) technology. The use of CAI technology allows for high quality collection of complex population-specific content sections. For example, the system facilitates the collection of the relationships of all household members to each other (i.e., the relationship grid). This wealth of information will enable a detailed analysis of family structures, an important concept for analysis. This type of collection would be very difficult to implement in a paper and pencil environment.

The CAI system has two main parts:

1) Case Management

The Case Management system controls the case assignment and data transmission for the survey. For this survey, a case refers to an individual selected for the LSIC sample. The Case Management system also automatically records management information for each contact (or attempted contact) with respondents and provides reports for the management of the collection process.

The Case Management system routes the questionnaire applications and sample file from headquarters to the regional offices and from the regional offices to the interviewers' laptops. The returning data takes the reverse route. To assure confidentiality, all data is encrypted before transmission. The data are unencrypted only once they are on a separate secure computer with no external access.

2) Survey-specific Components

Locating Respondents
The Wave 2 LSIC target population consists of immigrants who have been in Canada for only two years. For a variety of reasons, new immigrants are a highly mobile population during their first years in Canada. Respondent tracking is therefore necessary.

To help locate respondents a contact questionnaire was designed to request the immigrant's address in Canada (if known) as well as the address of a contact person in Canada. The form also contained a consent statement asking the respondent to grant Statistics Canada permission to access information held by other federal/provincial organizations, such as a provincial health department, for tracing purposes only. The form was enclosed in the packages provided to immigrants when they received their landing visa from a Canadian Mission Abroad.

Access to additional tracing information was only granted with consent from the potential respondent. This consent allowed Statistics Canada to obtain access to tracing-related information from health card records of all provincial health departments, with the exception of Nova Scotia. This source of information was considered to be the most current address information for the respondents.

Longitudinal Respondent Contact
In each wave, the first contact was established with the selected respondents using the address and telephone number provided on the sample file by Head Office. The interviewer confirmed that the respondent lived at that address. Once it was established that the interviewer was speaking to the correct person further steps were taken to ensure it was the proper respondent. Verification of respondent was done in two ways: matching of birth date and landing date.

Once the interviewer verified they had the proper respondent, the interviewer confirmed or updated the contact information (mailing and residence address, telephone number), as well as the list of household members. An appointment was then made to continue the interview in person.

If the interviewer was unable to locate the respondent the case was transferred to a designated tracing team in the regional offices, for further follow up.

Tracing Respondents
In each wave, within the regional offices, designated tracing teams followed up with further tracing sources to try and locate the respondent. Electronic phone books were the only effective public source used for tracing. The following sources of information were used for tracing the selected respondents:

- administrative files from Citizenship and Immigration Canada;
- survey contact questionnaires;
- addresses from provincial health cards (where an agreement with the province was reached and consent was given by the respondent); and
- electronic phone books (Québec, Ontario and British Colombia).

Person Most Knowledgeable
In the LSIC, proxy interviews are not allowed. The only exception is in the Income module, where the person most knowledgeable (PMK) regarding the families income is asked to answer these questions.

8.2 Collection

Collection Period
The survey uses a longitudinal design, meaning the same selected respondents are interviewed at different points in time. In LSIC, respondents are interviewed at three different points in time. The first of the three interviews is conducted six months after the respondent arrives in Canada; since it is desirable to assess their integration as soon as possible after they arrive. The second interview takes place two years after their arrival, and the final interview is conducted four years after their arrival.

To adequately represent the different immigration patterns in Canada over a one-year period, the sample is made up of 12 cohorts, consisting of 12 independent monthly samples selected over a period of 12 consecutive months.

Theoretically, an immigrant who arrived in October 2000 would be interviewed in April 2001, October 2002 and October 2004. In practice however, this may vary. Firstly, collection for the second wave began two months later than planned, in December 2002. Secondly, each monthly sample can remain in the field for up to three months for interviews to be conducted.

Landing date: October 2000 to September 2001		
Wave	**Collection Start**	**Collection End**
1	April 2001	May 2002
2	December 2002	December 2003
3	November 2004	November 2005

Collection for the First Two Waves
The collection of data for the first wave of the survey took place between April 2001 and May 2002 and collection for the second wave, between December 2002 and December 2003.

For Wave 1, most interviews (68%) were conducted in person, while the remaining interviews (32%) were conducted by telephone for various reasons (place of interview, specific language needs, etc.). In the second wave, just over half of the interviews were done in person.

Interviews were conducted in one of the 15 languages most frequently spoken by the target population: English, French, Chinese (Mandarin, Cantonese), Punjabi, Farsi/Dari (one language), Arabic, Spanish, Russian, Serbo-Croatian, Urdu, Korean, Tamil, Tagalog, and Gujarati. The 15 languages selected cover approximately 93% of the immigrant population in Canada.

Interview Length
On average in Wave 1, interviews lasted approximately 90 minutes. Fifteen minutes were devoted to the Entry and Exit components and the remaining 75 minutes to the survey. For Wave 2, interviews lasted approximately 65 minutes.

9.0 Data Processing

The main output of the Longitudinal Survey of Immigrants to Canada (LSIC) is a "clean" master data file. This chapter presents a brief summary of some of the processing steps involved in producing this file.

9.1 Initial Application Editing

Computer Generated Edits
As discussed earlier, all of the information for the sampled individuals was collected in a face-to-face, or telephone interview when a face-to-face was not possible, using a computer-assisted personal interviewing (CAPI) application. As such, it was possible to build various edits and checks into the questionnaire in order to ensure that high quality information was collected. Below are specific examples of the types of edits used in the LSIC computer-assisted interviewing (CAI) application:

Flow Pattern Edits
All flow patterns were automatically built into the CAI system. For example, for questions pertaining to a spouse/partner or child, the CAI system would automatically refer to the relationship information of all household members collected in the Entry Module to determine whether the longitudinal respondent (LR) had a spouse/partner or child living with them. If a spouse/partner or child was present, the CAI system continued with the specific questions related to them. If not, the CAI system automatically skipped these questions.

General Consistency Edits
Some consistency edits were included as part of the CAI system, and interviewers were able to "slide back" to previous questions to correct for inconsistencies. Instructions were displayed to interviewers for handling or correcting problems such as incomplete or incorrect data. For example, in the Language Module, if the respondent indicated that English was the language he/she most often spoke at home, the respondent could not answer that he/she do not speak English to a following question. If this happened, an edit screen popped up and the interviewer had to change one of the answers.

Range Edits in Numeric Fields
Range edits were also built into the CAI system for questions asking for numeric values. If numbers entered were outside the range, the system generated a pop-up window which stated the error and instructed the interviewer to make corrections to the appropriate question. For example, in the collection of the Employment Details sub-module, the number of hours worked per week was set to a maximum of 168 hours (the number of hours in a week). If the respondent indicated that he/she worked more than 168 hours a week, the range edit was triggered.

9.2 Minimum Completion Requirements

One of the first steps in processing the LSIC data consisted of determining the minimum number of responses required in order for a record to be considered valid.

No Information Collected
In some cases, no LSIC information was collected for a sampled individual. This happened when an interviewer was unable to trace a selected immigrant or was not able to make contact for the entire collection period. In other cases, the individual refused to participate in the survey, was away for the duration of the collection period or language barriers (an individual who did not speak one of the 15 survey languages) prevented an interview from taking place.

For cases where no information at all was collected for an immigrant, the individual was dropped from the LSIC file and the sampling weights for responding immigrants were inflated to account for these "dropped" immigrants.

Complete or Partial Response
Most of the time, respondents provided a complete response, meaning that all modules were completed. In some other cases, for various reasons, it was possible to conduct part of the interview. Some respondents only had a very limited amount of time to devote to the interview; or in some cases the interviewer did a portion of the interview with the respondent and arranged to complete it at another time but was unable to re-contact the respondent. Lastly, some respondents may have refused to answer one or more modules.

Criteria for Partial Response
For cases where the interview did not yield a complete response, it was necessary to have a set criteria for determining when a record was valid (partial response). A record was considered to be a partial response when the interview yielded enough information to apply imputation strategies to complete the remaining questions.

In the first wave, a record was considered to be a partial response when some modules were incomplete, with the exception of the first two, the Entry and Background modules. In the second wave, the criteria were slightly different. In order for a record to be valid, the respondent had to have at least provided answers to the Entry module. Partial response cases were retained in the sample of respondents.

Missing Components and Mass Imputation
For the partial responding individuals, all variables from the missing components were set to not stated or imputed, except for three modules: “Values and Attitudes,” “Citizenship” and “Perceptions of Settlement.” The questions in these modules asked about the LR’s personal opinions and perceptions, which vary too much to establish a solid mass imputation strategy. For more information on imputation, see Chapter 11.0.

Total Responding Records
In total, 9,322 longitudinal respondent records were determined to be complete enough to be kept in the final file for Wave 2.

These immigrants had resided in a total of 10,681 places prior to their current place of residence (collected in the Where Lived sub-module). They had taken a total of 8,170 courses or training sessions (collected in the sub-module on training programs), and reported 10,200 credentials of various kinds. They had a total of 11,959 jobs or businesses since landing in Canada (collected in the sub-module on jobs).

9.3 Coding

Three different levels of coding were done: open-ended questions, census type of questions, and text recorded in the “Other - Specify” fields. Given the number of new categories that were added to questions during the coding step, coding was done before the pre-edit step, in order to minimize adjustments to the pre-edit and flow edits stages.

9.3.1 Coding of Open-ended Questions

A few data items on the LSIC questionnaire were recorded by interviewers in an open-ended format. For example, in the Employment Module, a LR who had worked since they arrived in Canada was asked a series of open-ended questions about each job they have held:

- What kind of business, industry or service is/was it?
- What kind of work do/did you do in this job?
- In this job what are/were your most important duties?

In the Wave 1 questionnaire, in the Perceptions of Settlement Module, the last two questions were:

- What is the single most useful thing that was done to help you settle in Canada?
- What is the single most useful thing that could have been done to help you settle in Canada?

How they are recorded
The interviewer recorded, in words, the answer provided by the respondent to these questions. At Head Office, these written descriptions were converted into codes (e.g., industry or occupation) to make the data comparable.

How they are coded
The open-ended questions were coded using various standard classifications. Occupation questions were coded using the 1991 Standard Occupational Classification codes (SOC) and the industry questions were coded using the 1997 North American Industry Classification System (NAICS).

In the first wave, the variables asking about the major field of study in the Education module and the Studies sub-module were coded using the major field of study (MFS) code set. In the second wave, these same variables were coded using the Classification of Instructional Programs (CIP, Canada, 2000) code set. This is the classification system that Statistics Canada currently uses for the education field. To ensure the comparability of the data from the two waves, the Wave 1 variables were recoded using the CIP classification system.

Survey-specific code sets were developed in order to code questions such as the two examples from the Perceptions of Settlement Module.

9.3.2 Coding of Census Type Variables

A few of the LSIC questions were also asked in the 2001 Census. These include questions on country of birth, country of citizenship, language, religion, ethnic group and visible minority.

How they are recorded
For most of these questions, a pick-list was included in the questionnaire. In many cases, the “Other - Specify” category was chosen by interviewers and a text entry was recorded.

How they are coded
At Head Office, each of these questions were coded using the corresponding Census code set in order to match the 2001 Census data dictionary. The groupings resulting from the coding are then perfectly comparable with Census data.

9.3.3 Coding of “Other – Specify” Answers

In the LSIC questionnaire, several questions included an “Other - Specify” category, which allowed the interviewers to enter a text entry for an answer they could not find in the pick-list.

How they are coded
After a careful examination, items entered in the “Other – Specify” fields were coded according to three possible scenarios:

- the code for an existing category was assigned when the concept was similar;
- it remained in the “other” category;
- a new category was added to those originally included in the questionnaire, when the responses for a category represented approximately 5% or more of all responses.

9.4 Head Office Editing

Pre-edits
Before proceeding with the pre-edits, databases were created for the main section of the questionnaire, for the information collected on the LR’s household as well as for each of the roster files.

In the first pre-edit step, “Mark all that apply” questions were de-strung and values converted to Yes (1) or No (2) responses. Non-response values from the CAI system were also recoded to standard non-response codes for refusals, don't know and not stated.

Converting non-response codes to standard codes

Don’t know
During a CAI interview, the respondent may not know the answer to a particular item. The CAI system has a specific function key to press in such a situation.

In the LSIC files, the code used to indicate that the respondent did not know the answer to an item is "7". For a variable that is two digits long the code is "97", for a three-digit variable "997", etc.

Refusals
The respondent may choose to refuse to provide an answer for a particular item. The CAI system has a specific function key that the interviewer presses to indicate a refusal. This information is recorded for the specific item refused and transmitted back to Head Office.

In the LSIC files, an item which was refused is indicated by a code "8". For a variable that is two digits long the code is "98", for a three-digit variable "998", etc.

Not stated
In some cases, as part of Head Office processing, the answer to an item has been set to “not stated”. The not stated code indicates that the question was not asked of the respondent. These codes were assigned for three main reasons:

1) As part of the CAI interview, the interviewer was permitted to enter a “refusal” or “don't know” code, as described above. When this happened the CAI system was often programmed to skip out of this particular section of the questionnaire. In the case of refusal, it was assumed that the line of questioning was sensitive and it was likely that the respondent would not answer any more questions on this particular topic area. In the case of a “don't know” it was assumed that the respondent was not well enough informed to answer further questions and it was not known if the subsequent questions were applicable. As part of the LSIC processing system, it was decided that all of these subsequent questions should be assigned a “not stated” code.

2) In some cases, sections or entire modules of the questionnaire were not started or they were started but ended prematurely. For example, there may have been some kind of interruption, or the respondent decided that he/she wished to terminate the interview. If there was enough information collected to consider the module as responded, the questions that were not answered would be coded to “valid skip”. If an entire module was not answered, mass imputation was performed - with the exception of the Citizenship Module, Values and Attitudes Module and the Perceptions of Settlement Module, where questions not answered remained as “not stated”.

3) The third situation in which “not stated” codes were used was as a result of consistency edits. When the relationship between groups of variables was checked for consistency, if there was an error, often one or more of the variables were set to “not stated”.

In the case of derived variables, if one or more of the input variables contained a “not stated”, then the derived variable was also set to “not stated”.

An item which was coded as “not stated” is indicated by a code "9". For a variable that is two digits long the code is “99", for a three-digit variable "999", etc.

Flow edits and assignment of valid skip codes
As the last step of the pre-edits, the flow patterns for each of the files were processed and standard codes for “valid skips” were assigned (6, 96, and 996).

For example, for all questions where the LR did not have a spouse or common-law partner residing in the household, all "spouse" variables have been set to “valid skip”.

9.5 Consistency Edit

Consistency Editing
Consistency editing is carried out to verify the relationship between two or more variables. An example of a consistency problem could be when the personal income of the LR is higher than the total family income, of which it is only a part. We try to solve problems like this by using as much information as possible from other variables. If possible, we change the response to what seems to be the correct answer. But in some cases, the incoherent response was changed to “not stated”. As a result, in the final file, there are no remaining inconsistencies between the personal income of the LR and the total family income.

Relationship edits
Relationship edits are one form of consistency edits. For various reasons, relationship data collected in the Entry Module at times contained errors. The relationship edit step ensures a clean file and consistency in the relationships among members of the household.

For example some respondents whose spouse had children reported being “unrelated” to the children. In fact, according to the Census definitions, these people should have been step-parents, which is not a well-known concept for some recent immigrants to Canada. Similarly some foster parents reported being unrelated to a foster child, when they should have reported being foster parents.

9.6 Derived Variables

Utility of derived variables
Derived variables facilitate the work of analysts by providing condensed information for which the extraction requires a certain amount of programming. For example, a derived variable can be the result of the combination of answers from many different questions. Variables providing the count of events (like jobs for instance) present in a roster are another kind of derived variable.

Some derived variables were created so that Wave 2 data are comparable with data from Wave 1. In some cases, the responses to several questions have been combined to create a derived variable comparable (measuring the same item) to a variable that exists in Wave 1. For this reason, there are more derived variables in Wave 2.

Where to find the Derived Variables on the Files
With the exception of the Longitudinal Respondent's entity, which is mostly comprised of derived variables, the derived variables are usually placed after the questions in each entity to which they belong. Derived variables made from information from events rosters can be found in the entity to which it relates to (for example, the variable "number of places lived in before current place" is located in the Housing entity).

Derived Variable Name
All derived variables on the LSIC data files have a "d" as the fourth character of the variable name. For example, the name of the variable for the "Total hours per week currently in class or training" is ED1D008.

Some derived variables in the original Wave 1 file had to be renamed in the Wave 2 data file. In the first wave, the numbering of these variables started with 001. This caused a problem, since two variables could have the same item identifier, e.g., HS1Q001 and HS1D001. In Wave 2, the variable HS1D001 was renamed HS1D117 to avoid possible confusion. For further details, please refer to the concordance table:

With counts:
LSICWave2_ELICVague2\Concordance\LSIC_W2_Concordance_Table.pdf

Without counts:
LSICWave2_ELICVague2\Concordance\LSIC_W2_Concordance_Table_NoCnts.pdf

10.0 Non-response

A survey's response rates are a measure of the effectiveness of the population being sampled, the collection process and are also a good indicator of the quality of the estimates produced. As with other surveys, the Longitudinal Survey of Immigrants to Canada (LSIC) is faced with a certain level of non-response. This chapter will provide a summary that distinguishes between two types of non-response: total (or unit) non-response, and partial non-response.

Total non-response:
No information was collected for the sampled unit. This is the case of incomplete information as described in Section 9.2. For total non-response, some weighting adjustment methods were used to compensate. This topic is discussed in more detail in Chapter 12.0.

Partial non-response:
At least one but not all the modules were complete. The criteria for module completion are outlined in Section 9.2. Partial non-response was corrected by imputation.

10.1 Definition of Response Status

The following definitions outline the content of the tables below.

The **out-of-scope immigrant** in Wave 2 was an immigrant who was listed on sample file for Wave 2 but after some verification steps did not meet the criteria of the population of interest. Some examples of the out-of-scope are immigrants who, were deceased, or were institutionalized or moved outside Canada.

A **responding immigrant** in Wave 2 is the selected longitudinal respondent (LR) who responded in Wave 1 and is either a partial or complete respondent (see Section 9.2) in Wave 2. After Wave 2, 9,322 usable records were identified as responding units.

Unresolved or untraced refers to cases identified during Wave 2 collection where there was no contact at all with the selected immigrant. No information was collected as to their whereabouts.

Non-respondents refers to cases identified during Wave 2 collection where the selected immigrant was somehow located and confirmed to be in Canada, but for a given reason could not or chose not respond to the interview.

While both unresolved and non-respondent cases result in unusable records, the main difference between the two is that in cases of non-response the selected immigrant was confirmed to be in the Wave 2 population of interest.

Table 10.1 Results of Wave 2 Collection by Reference Month and Year

Month and Year	Respondents	Non-respondents	Out-of-scope	Unresolved	Total
October 2000	747	116	14	115	992
November 2000	800	99	11	131	1,041
December 2000	758	135	16	78	987
January 2001	747	121	14	90	972
February 2001	825	122	20	89	1,056
March 2001	783	128	15	63	989
April 2001	771	124	10	77	982
May 2001	823	94	20	111	1,048
June 2001	798	98	25	118	1,039
July 2001	794	116	10	92	1,012
August 2001	761	125	25	84	995
September 2001	715	92	20	100	927
Total	**9,322**	**1,370**	**200**	**1,148**	**12,040**

Reference month and year are the terms used to denote the month and year of landing

Table 10.2 Results of Wave 2 Collection by Class of Immigrant

Class of Immigrant	Respondents	Non-respondents	Out-of-scope	Unresolved	Total
Economic	5,412	745	132	684	6,973
Family	2,497	486	64	318	3,365
Refugees	1,317	128	3	142	1,590
Other	96	11	1	4	112
Total	**9,322**	**1,370**	**200**	**1,148**	**12,040**

Table 10.3 Results of Wave 2 Collection by Age Groups

Age Groups	Respondents	Non-respondents	Out-of-scope	Unresolved	Total
15 to 24	1,801	279	25	220	2,325
25 to 34	3,568	431	77	521	4,597
35 to 44	2,385	336	50	249	3,020
45 to 64	1,340	250	30	135	1,755
65 and over	228	74	18	23	343
Total	**9,322**	**1,370**	**200**	**1,148**	**12,040**

Table 10.4 Results of Wave 2 Collection by Sex

Sex	Respondents	Non-respondents	Out-of-scope	Unresolved	Total
Male	4,607	723	113	596	6,039
Female	4,715	647	87	552	6,001
Total	**9,322**	**1,370**	**200**	**1,148**	**12,040**

Table 10.5 Results of Wave 2 Collection by Intended Province of Destination

Province	Respondents	Non-respondents	Out-of-scope	Unresolved	Total
Newfoundland and Labrador	23	1	1	2	27
Prince Edward Island	9	0	0	0	9
Nova Scotia	49	3	1	10	63
New Brunswick	45	2	0	4	51
Quebec	1,341	159	30	178	1,708
Ontario	4,726	701	108	680	6,215
Manitoba	205	23	1	25	254
Saskatchewan	79	24	1	7	111
Alberta	1,058	167	20	58	1,303
British Columbia	1,787	290	38	184	2,299
Canada	**9,322**	**1,370**	**200**	**1,148**	**12,040**

Table 10.6 Results of Wave 2 Collection by Place of Birth

Place of Birth	Respondents	Non-respondents	Out-of-scope	Unresolved	Total
Africa	952	113	9	125	1,199
America	656	108	17	92	873
Asia	5,901	917	121	756	7,695
Europe	1,744	226	51	168	2,189
Oceania	69	6	2	7	84
Total	**9,322**	**1,370**	**200**	**1,148**	**12,040**

11.0 Imputation

Imputation is essentially the process by which a plausible value is used to replace a missing or inconsistent value. The goal is to construct values that will lead to approximately unbiased estimators. There are many well-known techniques available to impute values for a given record or variable. When carried out properly, imputation improves data quality by reducing non-response bias. In the Longitudinal Survey of Immigrants to Canada (LSIC), imputation was done to ensure that a complete data set of variables or records was produced and to minimize the "not stated" fields in the microdata file.

The next two sections include, respectively, a description of nearest-neighbour donor imputation used to address incomplete modules; and the techniques used for imputation of items in the Income Module.

11.1 Mass Imputation

11.1.1 Longitudinal Imputation

For Wave 2, the mass imputation strategy of Wave 1, as described in the LSIC Wave 1 User Guide, could have been repeated. But, by doing so, longitudinal inconsistencies could have been introduced. These inconsistencies would have arisen for a couple of reasons: either a given longitudinal respondent (LR) could be complete in one wave and partial in the other; or, for a partial LR in both waves, a different donor might be chosen by independent imputation. These inconsistencies are of particular concern when imputing roster data, as they are used in the derivation of other variables. A roster is a data file with as many records for a given LR as the number of events for a concept of interest, such as employment history.

In order to overcome these limitations and to save potential processing time a longitudinal mass imputation technique was established. The mass imputation at Wave 2 was longitudinal in the sense that imputation was done simultaneously for data collected at both waves.

The first step was to identify which modules had to be imputed longitudinally. For this purpose longitudinal completion codes were generated. As discussed in Section 9.2, *keyfields* were defined for Wave 2 on the same principles as in Wave 1. Based on Wave 1 and Wave 2 completion codes, a longitudinal response code was established. A Wave 2 LR was deemed as a longitudinal complete respondent if and only if the LR was a complete respondent in both waves. Otherwise the LR was considered as a longitudinal partial respondent. A consequence of this rule was the classification of a module as longitudinally incomplete if the module was incomplete in either wave. Thus in instances where a module was complete in one wave but not in the other, legitimate data for the particular module were overwritten for one wave. Fortunately there was a small number (552 out of 9,322) of LRs for whom that was an issue.

Table 11.1 presents the different patterns of longitudinal module completion for all responding records. In the table, a "1" denotes that the module is complete, i.e. all *keyfields* within the module (Wave 1 and Wave 2) have valid values, while a "2" indicates that the module is incomplete (information is incomplete for one or both waves).

Table 11.1 Distribution of Longitudinal Module Completion

EN	SI	LS	HS	ED	EM	HL	IN	Number of Records	Percent
2	2	2	2	1	1	1	1	1	0.01%
2	1	2	1	2	1	1	1	1	0.01%
2	1	1	1	1	2	2	2	4	0.04%
2	1	1	1	1	1	2	2	3	0.03%
2	1	1	1	1	1	1	2	6	0.06%
2	1	1	1	1	1	1	1	2	0.02%
1	2	2	2	2	2	2	2	17	0.18%
1	2	1	2	2	2	2	2	1	0.01%
1	2	1	2	1	1	2	2	1	0.01%
1	2	1	1	1	1	2	2	1	0.01%
1	2	1	1	1	1	2	1	1	0.01%
1	2	1	1	1	1	1	2	8	0.09%
1	2	1	1	1	1	1	1	28	0.30%
1	1	2	2	2	2	2	2	8	0.09%
1	1	2	1	2	1	1	1	1	0.01%
1	1	2	1	1	2	2	2	1	0.01%
1	1	2	1	1	1	1	1	13	0.14%
1	1	1	2	2	2	2	2	9	0.10%
1	1	1	2	2	1	1	2	1	0.01%
1	1	1	2	1	1	1	2	6	0.06%
1	1	1	2	1	1	1	1	6	0.06%
1	1	1	1	2	2	2	2	15	0.16%
1	1	1	1	2	1	1	2	4	0.04%
1	1	1	1	2	1	1	1	20	0.21%
1	1	1	1	1	2	2	2	40	0.43%
1	1	1	1	1	2	1	2	2	0.02%
1	1	1	1	1	1	2	2	36	0.39%
1	1	1	1	1	1	2	1	15	0.16%
1	1	1	1	1	1	1	2	327	3.51%
1	1	1	1	1	1	1	1	8,744	93.80%

EN – Entry; SI - Social Interaction; LS – Language Skills; HS - Housing; ED - Education; EM - Employment; HL - Health; IN - Income.

Table 11.1 shows that the Income Module was the least reported module longitudinally with 5% non-response. For the Income Module, a different processing approach was used. This approach is described in Section 11.2.

11.1.2 Strategy for Longitudinal Imputation

For longitudinal partial non-response in Wave 2, mass imputation for the incomplete modules was carried out using the nearest-neighbour donor technique. The donor imputation method generally would not alter the distribution of the data, which is a

drawback of many other imputation techniques. It aimed at replacing missing information for a longitudinal partial respondent with values provided by a longitudinally complete respondent who is “similar” to him/her. It worked in the following manner: based on a statistical distance calculated on selected socio-demographic information, a donor (longitudinal complete respondent) determined to be the closest to the recipient (longitudinal partial respondent) was identified and the values of the donor were used to replace the missing values for the recipient in both waves. This was conducted module by module. It is worth noting that the socio-demographic variables used in donor selection included the variables that determined the questionnaire skip-pattern: the presence of LR’s spouse and children, and also the presence of LR’s school-age children.

For a longitudinal partial respondent for whom more than one module was incomplete, the same donor record was used for all the incomplete modules. Note that only complete and edited records were used as potential donors. To keep consistency within variables, the complete set of variables for a given module of the donor was imputed into the recipient record. At the end of this process, all records had fully completed modules. A flag indicating whether a module was imputed was created.

11.1.3 Imputation for Events

Another aspect of mass imputation in Wave 1 was the adjustment of different date variables from the rosters (housing, education and training, employment history). The dates were first imputed using the donor record and then adjusted so that they would be consistent with the recipient’s landing and interview dates. The adjustment was done with respect to the interview date. This method had the important drawback to alter the time between the landing date and the dates at which the different events occurred. When imputed dated happened to be earlier than landing dates, these dates had to be modified. The result was that the distribution of some variables based on landing date could have altered e.g.: time elapsed from landing to the first job. Also, by adjusting imputed dates, there was a potential for the recipient LR to have seasonal employment in the wrong season, e.g. snow removal in July! At the end, the imputed adjusted dates neither belonged to the donor nor the recipient.

After considering the deficiencies of adjusting imputed dates in Wave 1, it was not repeated for Wave 2 longitudinal mass imputation. The imputed dates of the recipient were simply the donor dates. The donor’s interview dates for both waves, landing date and the number of days between landing date and interview dates are provided on the recipient records and were used to derive related variables. The imputed data paints a picture for the recipient had they arrived and been interviewed at the same time as their donors.

Additional variables related to donor’s intended destination at landing, geographical information regarding moving history inside Canada, and trade or occupation practised or intended to practice at landing in Canada are also provided on the recipient records. It is possible for an imputed LR to have: an out-of-province move when in fact no or local move had happened; or, a complete change in occupations before and after imputation. Thus donor’s geographical information is useful in mapping residency and moving pattern inside Canada for the recipients based on donor’s data. Similarly, the donor’s occupation at landing is important in establishing the continuity or changes of occupations held by the recipient over time.

It should be kept in mind that for the recipients the imputed data corresponds to the time frame, location, and characteristics of the donors. One simply should not compare the imputed data of the recipients with their actual data especially for roster data.

11.2 Field Imputation for Income Variables

The immigrant interview in the LSIC includes a number of questions on income. Information is collected on the longitudinal respondent's family income from different sources within Canada and outside Canada. Information is also collected on the longitudinal respondent's personal income from all sources (within and outside Canada) and on the amount of his/her savings and loans.

Income is a sensitive topic. Some respondents refuse to give answers to the detailed questions on the various sources of income. Among such respondents, some nevertheless provide an estimate of total family income or an estimate of their personal income, sometimes using income intervals. As well, among those who answer the questions, it may happen that the amounts indicated in the sections concerning income are incompatible with the answers given in the section concerning employment (for example, according to the answers given in the employment section, the respondent worked during the last 12 months but reports no wages and no net income from self-employment in the section on income). Income is then imputed to fill in missing values attributable to partial non-response (Section 11.2.2) and, to a lesser extent, to correct inconsistent data where possible (Section 11.2.1).

11.2.1 Detection and Imputation of Outliers

Before field imputation of missing values is carried out, quantitative income variables first go through an outlier detection process. One of the purposes of this process is to define the donor pool that will be used to impute the missing values. For each quantitative variable, the weighted empirical distribution is produced and graphically represented in order to compare the data obtained and identify extreme values. It should be noted that income data are generally asymmetrical. Their asymmetry is characterized by a larger spread toward the high values of the variable and the fact that some data can take on negative values (e.g., negative income in the case of self-employment). Values identified as extreme are inspected manually. The inspection can give rise to two possible results:

1) The value is an outlier: in this case, the median or a value more plausible than the median[6] is imputed;
2) The value is extreme but acceptable in light of other information: in this case, the value is not changed but is identified for exclusion from the pool of donors for imputation.

11.2.2 Field Imputation of Missing Values

Missing values in the income module are then imputed by the nearest neighbour method. This method consists in locating a respondent who provided a response to the income section (a donor) and whose characteristics are similar to those of the person or family that did not provide complete information on income (a recipient). Once the nearest neighbour has been identified, the amount reported by the donor is imputed to the recipient. Since the rules for finding a donor differ depending on the income source to be imputed, the imputation is done by field (that is, independently for each source). In other words, in a case where more than one income source had to be imputed, there might be more than one donor.

The data file released for Wave 2 is a longitudinal file, meaning that it contains both Wave 1 and Wave 2 data for Wave 2 respondents. Mass imputation was thus carried out longitudinally to ensure consistency between Wave 1 data and Wave 2 data (see Section

[6]. One situation in which a value more plausible than the median is imputed is where a capture error has occurred, e.g., 200,000 was entered instead of 20,000. In this case, 20,000 will be the imputed value.

11.1). This had the effect of changing some Wave 1 data, including data in the income module. The field imputation process for income variables must then be carried out for both Wave 1 data and Wave 2 data. The two imputation processes are carried out independently. However, for the imputation of Wave 1 variables, the donor pool is limited to immigrants who were also respondents in Wave 2. While Wave 1 respondents who did not respond in Wave 2 might technically serve as donors for Wave 1 data, these individuals might have different characteristics from immigrants who responded in both waves. They are therefore excluded from the donor pool to avoid introducing a potential bias in the data.

In the LSIC, only amounts of family income from 11 sources within Canada are imputed, in addition to the longitudinal respondent's personal income. Among the variables that represent income sources within Canada, six are related to the labour market and five are transfer payments, meaning income from a government in Canada. The list of variables for which imputation was carried out is given in Table 11.2. The table shows the overall imputation rate for each variable, for Wave 1 and Wave 2 respectively. It should be pointed out that even though imputation generally improves data quality overall, the artificial data created are used for estimation purposes and can lead to a substantial underestimation of variance, especially if the imputation rate is high. Imputation flags are integrated into the LSIC file to identify variables for which there was an imputation in a record. Users can thus measure the scope of imputation for a particular variable. For all imputation flags in the LSIC data file, a "I" appears as the fourth character of the variable name. Thus, IN2I004 is the imputation flag for family income from all jobs (IN2Q003).

Table 11.2 Imputation Rates for Income and Earnings

Variable Description	Wave	Variable Name	Name of Imputation Flag for Variable	Number of Cases Excluding Valid Skips	Number of Imputed Values	Imputation Rrate
Income from all jobs	Wave 1	IN1Q003	IN1I004	6,141	759	12.36%
	Wave 2	IN2D003x	IN2I004	7,565	744	9.83%
Income from self-employment	Wave 1	IN1Q005	IN1I006	360	102	28.33%
	Wave 2	IN2D005x	IN2I006	1,162	260	22.38%
Canadian business or company pension	Wave 1	IN1Q027	IN1I028	35	13	37.14%
	Wave 2	IN2D027x	IN2I028	42	1	2.38%
Private sponsor	Wave 1	IN1Q030	IN1I031	67	4	5.97%
	Wave 2	IN2D030x	IN2I031	72	7	9.72%
Investments	Wave 1	IN1Q033	IN1I034	293	44	15.02%
	Wave 2	IN2D033x	IN2I034	384	25	6.51%
Other sources	Wave 1	IN1Q036	IN1I037	454	24	5.29%
	Wave 2	IN2D036x	IN2I037	494	14	2.83%
Social assistance	Wave 1	IN1Q008	IN1I009	1,267	28	2.21%
	Wave 2	IN2D008x	IN2I009	1,245	29	2.33%
Employment insurance	Wave 1	IN1Q011	IN1I012	242	35	14.46%
	Wave 2	IN2D011x	IN2I012	1,395	81	5.81%
Child tax benefit or credits	Wave 1	IN1Q014	IN1I015	3,007	161	5.35%
	Wave 2	IN2D014x	IN2I015	4,941	301	6.09%
Canada or Quebec pension plan	Wave 1	IN1Q017	IN1I018	136	22	16.18%
	Wave 2	IN2D017x	IN2I018	145	15	10.34%
Other government sources	Wave 1	IN1Q023	IN1I024	673	41	6.09%
	Wave 2	IN2D023x	IN2I024	1,029	27	2.62%
Longitudinal respondent's personal income from all sources	Wave 1	IN1D067	IN1I068	9,322	258	2.77%
	Wave 2	IN2D067x	IN2I068	9,322	370	3.97%

12.0 Treatment of Total Non-response and Weighting

The Longitudinal Survey of Immigrants to Canada (LSIC) is a probability survey. As is the case with any probability survey, the sample is selected to represent a reference population - the immigrant population - at a specific date within the context of the survey as accurately as possible. Each unit in the sample must therefore represent a certain number of units in the population. The complete sample for Wave 2 is a subset of the Wave 1 sample, consisting solely of immigrants responding in Wave 1. While this chapter makes some links between Wave 1 and Wave 2, it mainly deals with the weighting of Wave 2. For further details on the weighting of Wave 1, see Chapter 10.0 of the Wave 1 User Guide.

12.1 Representativity of the Weights

For most surveys, the sum of the final weights represents the estimated target population counts which usually equate to the population of interest. However, in the case of the LSIC, because of the mobility of the population and the survey objectives (see Chapter 3.0 from the Wave 1 User Guide), the population of interest is actually a portion of the target population, namely immigrants who were still residing in Canada at the time of interview. Furthermore, the Wave 2 population of interest differs from the Wave 1 population of interest since it is merely a subset of the latter.

Recall that the survey frame covers the target population - immigrants who meet all of the following criteria:

- arrived in Canada between October 1, 2000 and September 30, 2001;
- were age 15 or older at the time of landing;
- landed from abroad, must have applied through a Canadian Mission Abroad.

However, some of these immigrants resided in Canada for awhile before returning to their original country or migrating to another country. These immigrants do not have similar adaptation characteristics as the ones who are permanently residing in Canada. It is biased to include in the same weight adjustment the immigrants who moved out of Canada and those who still reside in Canada. The target population includes these two basic sub-groups.

The **Wave 2 population of interest** (PI**)** consists of immigrants from the LSIC who are still in Canada two years after their arrival (by comparison, the Wave 1 population of interest consisted of immigrants from the LSIC who were still in Canada six months after their arrival). The Wave 2 final weight yields unbiased estimates of the Wave 2 population of interest. The **out-of-interest population (**OOI) consists of immigrants who no longer live in Canada, i.e., who have left since landing in Canada.

12.2 Overview of the Weight Adjustments

During collection, there were four possible classifications for a selected immigrant; respondent, non-respondent, not in the population of interest, and unresolved. The first three categories resulted in an initial contact with the immigrant or with someone who was able to confirm their status. These cases are defined as resolved cases as the immigrant has a known status. The last collection outcome is the unresolved cases. For these, no contact was established and they remained unresolved. No information on whether they were still in Canada was available. The weight adjustments reflect these outcomes.

The sample can first be split between the resolved and the unresolved cases:

Sample $S = S_U + S_R$

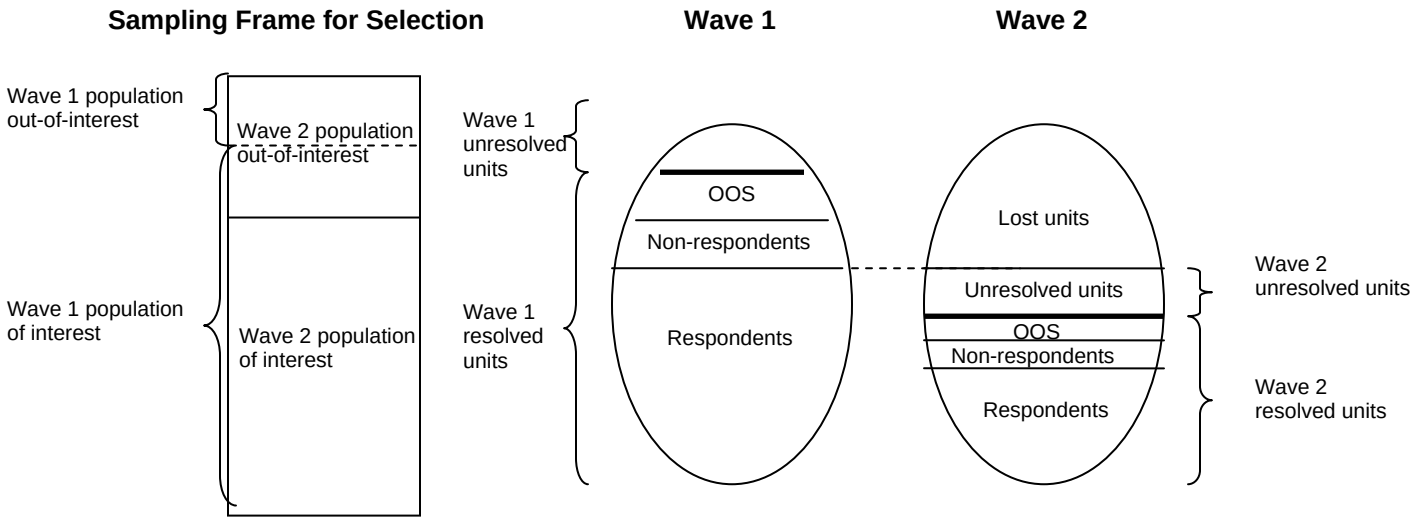
where S_U = sampled units unresolved
S_R = sampled units resolved

Furthermore, in the resolved portion $S_R = S_{RR} + S_{RN} + S_{RO}$

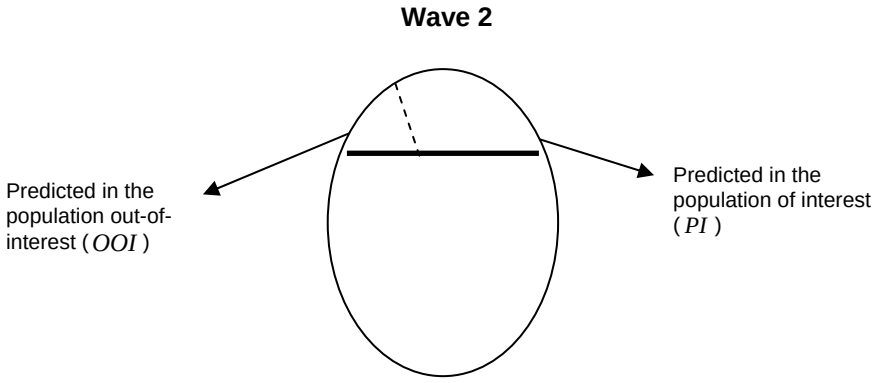
where S_{RR} = sampled units resolved that are respondents
S_{RN} = sampled units resolved that are non-respondents
S_{RO} = sampled units resolved that are not in the population of interest (referred to as OOI, out-of-interest).

Individuals who are out-of-scope are represented by OOS.

The following diagram presents an overview of these concepts as they relate to weighting and shows the passage from the sampling frame to the Wave 1 sample and then to the Wave 2 sample.



Conceptually, for the set of units that remained unresolved (S_U), it is fair to assume it is composed of units in the population of interest (PI) and in the population out-of-interest (OOI). However, at that point of the process, there was no information available. Consequently, the first step of the weighting process was to predict for the unresolved units whether they would have been in the population of interest or not. Through models, using the information available on the frame, information collected in Wave 1 and information on the resolved units in Wave 2, the status of the unresolved units was predicted as: PI or OOI as shown in the following diagram.



After this first step, we have a status (predicted or confirmed) for each selected unit indicating if they were part of the population of interest or not in the population of interest. Note that in the resolved units, the population of interest is composed of respondents and non-respondents. Thus the following notation will be used in subsequent sections:

For the unresolved units (s_U):

$j \in S_U$ = unresolved sampled units where $S_U = \hat{S}_{U\_PI} + \hat{S}_{U\_OOI}$

$j \in \hat{S}_{U\_PI}$ = unresolved sampled units predicted as PI

$j \in \hat{S}_{U\_OOI}$ = unresolved sampled units predicted as OOI

For the resolved units (s_R):

$i \in S_R$ = resolved sampled units where $S_R = S_{RR} + S_{RN} + S_{RO}$

$i \in S_{RN}$ = resolved non-respondents units

$i \in S_{RR}$ = resolved respondents units

$i \in S_{RO}$ = resolved OOI units

12.3 Longitudinal Weighting for Responding Immigrants

The LSIC weighting strategy is based on a series of cascading adjustments. The final longitudinal weight is obtained by applying various adjustments to the initial weight. There are four weights involved in the weighting process which will compose the final weight; the initial weight, the non-response adjustment weight, the unresolved adjustment weight and finally the post-stratification weight. Table 12.1 shows the relationship between the different categories of outcomes related to the adjustment.

Table 12.1 Process of Classifying the Respondents Outcome Status

<table>
<tr><th>Sample</th><th>Tracing</th><th>Status</th><th colspan="2">Response</th></tr>
<tr><td rowspan="6">Wave 1 responding units</td><td rowspan="5">Resolved units</td><td rowspan="5">PI :
In scope units</td><td colspan="2">Responding units</td></tr>
<tr><td rowspan="4">Non-responding units</td><td>Refusal</td></tr>
<tr><td>Language problems</td></tr>
<tr><td>LR absent</td></tr>
<tr><td>Other non-response</td></tr>
<tr><td colspan="3">OOI (Left Canada, dead, etc)</td></tr>
<tr><td colspan="4">Unresolved units</td></tr>
</table>

Note that on the microdata file, only the responding resolved units, ($i \in S_{RR}$), have a final weight as they are the only units which have fully completed records. As for the out-of-interest population, ($i \in S_{RO}$) they also have a final weight, but are not available on the microdata file as they do not have full records. Only tabulations of this sub-population using the final weights are available.

The subsequent sections describe the initial weights (Section 12.3.1), the two weight adjustments, i.e. for non-response and unresolved units (Section 12.3.2) and finally post-stratification is explained in Section 12.3.3.

12.3.1 Initial Weight

At the time of selection, an initial design weight was assigned to the selected person. It is simply the inverse of the probability of selection of immigrants, and that probability depends on the selection method. Since a two-stage sampling method was used for the LSIC, the sampling weight attributed to each person selected is equal to the inverse of the probability of selection of the immigrant unit to which the person belonged, multiplied by the number of eligible persons in this immigrant unit.

For the Wave 1 weighting, the initial weight was the sampling weight described above. This weight was then adjusted to take non-response and unresolved cases into account. Lastly, a post-stratification adjustment was applied to achieve consistency with updated population figures. For more details on the sampling weight and the different Wave 1 adjustments, see Chapter 10.0 of the Wave 1 User Guide.

For the Wave 2 weighting, the initial weight is the Wave 1 weight before post-stratification, that is, the sampling weight adjusted for Wave 1 non-response and unresolved cases, and it is formulated as follows:

$$Initial\ weight = (sampling\ weight) * (non-response\ adjustment_{Wave\,1}) * (unresolved\ adjustment_{Wave\,1})$$

Algebraically, the initial weight for the Wave 2 weighting is:

$$w_{initial} = w_D * \left[\frac{\sum_{G_1^{(1)}} \sum_{i \in S_{RR}} w_D + \sum_{G_1^{(1)}} \sum_{i \in S_{RN}} w_D}{\sum_{G_1^{(1)}} \sum_{i \in S_{RR}} w_D} \right] * \left[\frac{\sum_{G_2^{(1)}} \sum_{j \in \hat{S}_{U\_PI}} w_D + \sum_{G_2^{(1)}} \sum_{i \in S_{R\_PI}} w_1}{\sum_{G_2^{(1)}} \sum_{i \in S_{R\_PI}} w_1} \right]$$

where $w_1 = w_D * \left[\dfrac{\sum_{G_1^{(1)}} \sum_{i \in S_{RR}} w_D + \sum_{G_1^{(1)}} \sum_{i \in S_{RN}} w_D}{\sum_{G_1^{(1)}} \sum_{i \in S_{RR}} w_D} \right]$

where $G_1^{(1)}$ = Wave 1 non-response adjustment class

$G_2^{(1)}$ = Wave 1 unresolved adjustment class

$w_{initial}$ = initial weight for Wave 2

w_D = sampling weight (for further details, see Section 10.3 of the Wave 1 User Guide)

12.3.2 Non-response and Unresolved Cases Weight Adjustments

For the Wave 2 resolved responding units ($i \in S_{RR}$), the weight adjustment has the following formulation [before the post-stratification adjustment]:

$$Intermediate\ weight = (initial\ weight) * (non-response\ adjustment) * (unresolved\ adjustment)$$

or

$$= initial\ weight * \frac{weighted\ \ sum\ of\ resolved\ \ units\ \ (respondent\ \ s\ and\ \ non-respondent\ \ s)}{weighted\ \ sum\ of\ respondent\ \ s} * \frac{weighted\ \ sum\ of\ resolved\ \ units\ and\ \ PI\ resolved\ \ units\ \ per\ \ prediction}{weighted\ \ sum\ of\ resolved\ \ units}$$

or algebraically

$$w_{\text{int}\_PI} = w_{initial} * \left[\frac{\sum_{G_1^{(2)}} \sum_{i \in S_{RR}} w_{initial} + \sum_{G_1^{(2)}} \sum_{i \in S_{RN}} w_{initial}}{\sum_{G_1^{(2)}} \sum_{i \in S_{RR}} w_{initial}} \right] * \left[\frac{\sum_{G_2^{(2)}} \sum_{j \in \hat{S}_{U\_PI}} w_{initial} + \sum_{G_2^{(2)}} \sum_{S_{R\_PI}} w_1}{\sum_{G_2^{(2)}} \sum_{i \in S_{R\_PI}} w_1} \right]$$

where $w_1 = w_{initial} * \left[\dfrac{\sum_{G_1^{(2)}} \sum_{i \in S_{RR}} w_{initial} + \sum_{G_1^{(2)}} \sum_{i \in S_{RN}} w_{initial}}{\sum_{G_1^{(2)}} \sum_{i \in S_{RR}} w_{initial}} \right]$

where $G_1^{(2)}$ = Wave 2 non-response adjustment class

$G_2^{(2)}$ = Wave 2 unresolved adjustment class

W_{int_PI} = intermediary weight of the population of interest PI

$W_{initial}$ = initial weight

Note: Section 12.3.4 discusses in greater detail the concept of adjustment classes for non-response and unresolved cases.

For the Wave 2 resolved out-of-interest population ($i \in S_{RO}$), there is only one adjustment, i.e., one adjustment to compensate for the predicted out-of-interest ($j \in \hat{S}_{U\_OOI}$) in the unresolved one.

$$w_{\text{int}\_OOI} = w_{initial} * \left[\frac{\sum_{G_2^{(2)}} \sum_{j \in \hat{S}_{U\_OOI}} w_{initial} + \sum_{G_2^{(2)}} \sum_{i \in S_{RO}} w_{initial}}{\sum_{G_2^{(2)}} \sum_{i \in S_{RO}} w_{initial}} \right]$$

12.3.3 Post-stratification

The purpose of post-stratification is to ensure consistency between the estimates produced from the survey and population estimates produced by an independent external source. Since the LSIC Wave 2 final weights give estimates of the Wave 2 population of interest and not the target population (see Section 12.1 on ***Representativity of the Weights***) and since there is no independent external administrative source on this subject, the post-stratification totals must be estimated. Since for Wave 1, a post-stratification file was available (in other words, immigrant population sizes in the post-strata were known from an external source), the post-stratification totals for Wave 2 can be estimated as follows:

$$\hat{N}_k^{(2)} = N_k^{(1)} - \sum_{i \in k \cap OOI} W_{fi}^{(1)}$$
$$= \sum_{i \in k \cap PI} W_{fi}^{(1)}$$

where $\hat{N}_k^{(2)}$ = estimated size of the immigrant population (PI) in post-stratum k (post-stratification total of post-stratum k for Wave 2)

$N_k^{(1)}$ = known size of the immigrant population in post-stratum k (post-stratification total of post-stratum k for Wave 1)

$W_{fi}^{(1)}$ = Wave 1 final weight of immigrant i

For the Wave 2 sample, the population of interest consists of all immigrants in the LSIC who are still in Canada two years after their arrival. Consequently, the post-stratification adjustment for this sample ensures consistency between the sum of the weights and the demographic estimate associated with this period for each combination of age, sex, place of birth (aggregated by region of the world) and class of immigrant. Tables 12.2 through 12.5 provide the detailed categories.

Table 12.2 Age Groups

15 to 24
25 to 34
35 to 44
45 and over

Table 12.3 Sex

Male
Female

Table 12.4 Place of Birth

Region	**World Area (WA)**
Central Africa	1 - Africa
Eastern Africa	
Northern Africa	
Southern Africa	
Western Africa	
Central America	2 - America
Northern America	
Southern America	
Caribbean and Bermuda	
Eastern Asia	3 - Asia
Southeast Asia	
Southern Asia	
West Central Asia and Middle East	
Eastern Europe	4 - Europe
Northern Europe	
Southern Europe	
Western Europe	
Oceania	5 - Oceania

Table 12.5 Immigrant Classes

Family Class
Economic Class – Skilled Workers (Principal Applicant)
Economic Class – Skilled Workers (Spouse and Dependents)
Economic Class – Business Independent and Other Independent Immigrants
Government Sponsored Refugees
Other Refugees

The variables are cross-tabulated except in the following situations:

- For Oceania, there is only one other cross-tabulation: Family versus all other immigration classes collapsed together. There is neither sex nor age grouping for the post-stratification.
- For Government Sponsored Refugees the age groups 35 to 44 years and 45 years and over are collapsed.
- For Other Refugees, there is neither sex nor age grouping for the post-stratification.
- For Economic Class – Business Independent and Other Independent, there is no sex grouping, and the age groups 25 to 34 and 35 to 44 are collapsed.
- For Family Class of immigrants from Africa, age 35 to 44 years, sex was collapsed.
- For Economic Class – Business Independent and Other Independent from the Americas, there is no age grouping for the post-stratification.
- For Economic Class – Business Independent and Other Independent from Africa, the age groups 25 to 44 and 45 and over are collapsed (Wave 2 only).
- For Economic Class – Skilled Workers (Principal Applicant) from the Americas, the age groups 15 to 24 and 25 to 34 are collapsed for males and females and the age groups 35 to 44 and 45 and over are collapsed for females only (Wave 2 only).
- For Economic Class – Skilled Workers (Principal Applicant) from Asia, the age groups 15 to 24 and 25 to 34 are collapsed for males and females (Wave 2 only).
- For Economic Class – Skilled Workers (Principal Applicant) from Europe, the age groups 15 to 24 and 25 to 34 and the age groups 35 to 44 and 45 and over are collapsed for females only (Wave 2 only).
- For Economic Class – Skilled Workers (Spouse and Dependants) from Africa, the age groups 35 to 44 and 45 and over are collapsed for males and females (Wave 2 only).
- For Economic Class – Skilled Workers (Spouse and Dependants) from the Americas, the age groups 35 to 44 and 45 and over are collapsed for males and females (Wave 2 only).

The adjustment has the following form:

$$\textit{Final weight} = \textit{Intermediate weight} * \frac{\textit{Estimated size of immigrant population in the population of interest (PI)}}{\textit{Estimate of population figures using intermediate weights}}$$

or algebraically for $i \in S_{RR}$,

$$W_f = \sum_{i \in S_{RR}} w_{\text{int\_}PI} * \frac{\hat{N}_k^{(2)}}{\sum_k \sum_{i \in S_{RR}} w_{\text{int\_}PI} + \sum_k \sum_{i \in S_{RO}} w_{\text{int\_}OOI}}$$

12.3.4 Adjustment Classes: Homogeneous Groups

The weight adjustment classes, as well as the post-stratification groups, are constructed under the same assumption. They must be homogeneous groups related to the correction being made. The non-response adjustment classes are constructed based on the homogeneity of responses within a class, meaning that they have the same probability of response. The unresolved adjustment classes were constructed based on homogeneity or a similar propensity of being resolved and being in scope.

For the LSIC, the non-response and the unresolved adjustment classes were derived using logistic regression models predicting respectively, the response probability and the resolution probability. For the latter model, the explanatory variables for predicting the population of interest status were included by default in the model.

The predictors or explanatory variables for the model predicting **responses** were; **age group**, **status of respondent during immigration process**, **country of citizenship**, **the indicator of non-response for the Education entity in Wave 1** and **the indicator of non-response to the question in Wave 1 asking the total employment income received since arrival**.

The explanatory variables for the model predicting the propensity of being **resolved** were **the number of immigrants in the immigration unit, the Wave 1 collection period** and **the province of destination in Canada**. In this model, the predictor of being **in the population of interest, the indicator of having visited Canada as a tourist before immigrating to Canada**, **the number of years of education** and **age group** were included by default. The classes were constructed using similar probabilities obtained from each respective model. The number of classes for each adjustment was defined based on a convergence algorithm ensuring unbiased estimates.

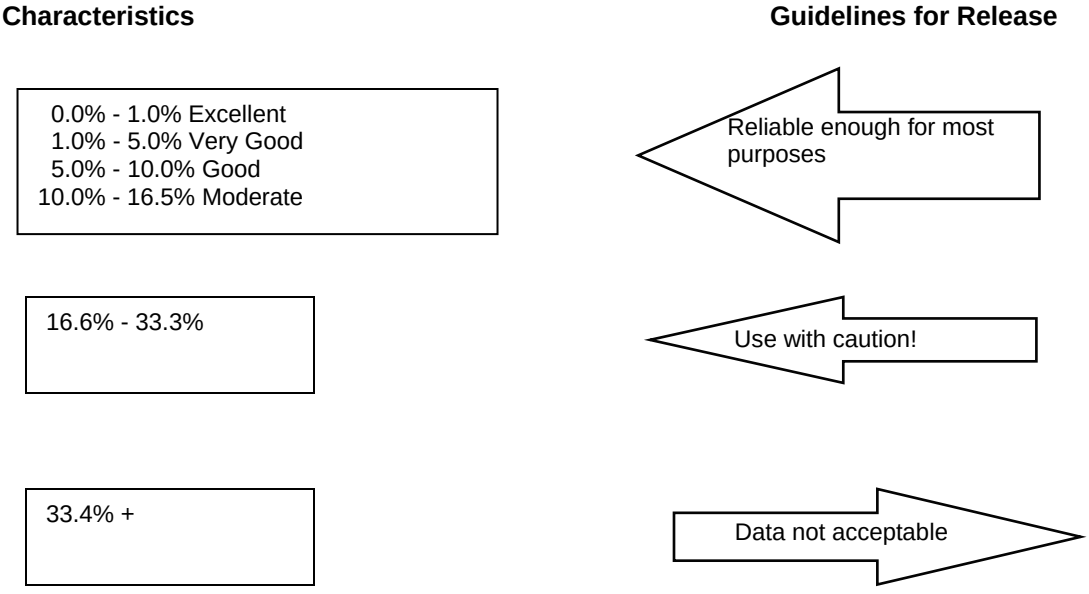
13.0 Data Quality and Coverage

This chapter provides the user with information about the various factors affecting the quality of the survey data. There are two main types of errors: sampling errors and non-sampling errors. A sampling error is the difference between an estimate derived from a sample and the one that would have been obtained from a census that used the same procedures to collect data from every person in the population. All other types of errors such as frame coverage, response, processing and non-response are non-sampling errors. Many of these errors are difficult to identify and quantify. These are discussed in Section 13.2.

13.1 Sampling Errors

The estimates derived from this survey are based on a sample of immigrants and not from a complete enumeration (census) under similar conditions. This difference is the sampling error of the estimates. Statistics Canada's *Standards and Guidelines on the Documentation of Data Quality and Methodology*[7] states that external users must be given an indication of the magnitude of the sampling error. It is highly recommended that users analyzing data or producing estimates from the Longitudinal Survey of Immigrants to Canada (LSIC) data files also provide their audience with indicators of the data quality.

The basis for measuring sampling error is the standard error of the estimates, estimated from the survey results. However, because of the large variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed relative to the estimate to which it pertains. This measure, known as the coefficient of variation (CV) of an estimate, is obtained by expressing the standard error of the estimate as a percentage of the estimate. The smaller the CV, the smaller the sampling variability, meaning smaller CVs are more desirable. The CV depends on the size of the sample on which the estimate is based, the population size and on the distribution of the sample, i.e. the sampling fraction of the units of the domains being estimated. The following diagram presents the characteristics of some coefficients of variation and the Statistics Canada guidelines for release.



[7] Statistics Canada. *Standards and Guidelines on the Documentation of Data Quality and Methodology*, 2002, www.statcan.ca/english/about/policy/infousers.htm.

13.2 Non-sampling Errors

There are many sources of non-sampling errors that are not related to sampling, but may occur at almost any phase of a survey operation. Interviewers may misunderstand survey instructions, respondents may make a mistake in answering the questions, responses may be recorded in the questionnaire incorrectly or errors may be made in the processing or tabulating of the data. For the LSIC, quality assurance measures were implemented at each phase of the data collection and processing cycles to monitor the quality of the data. These measures included precise interviewer training with respect to the survey procedures and questionnaire, observation of interviews to detect questionnaire design problems or misinterpretation of instructions, monitoring of final coding, and coding and edit quality checks to verify the processing logic. Chapter 9.0 outlines data processing procedures. Other kinds of non-sampling error are more easily quantifiable, especially non-response and the population frame coverage, the topics of the next two sections.

13.3 Non-response and Unresolved Cases

Non-response and unresolved cases, if not appropriately corrected, are the types of error that can lead to bias in the survey estimates. For the LSIC, these two types of response categories reduced significantly the number of usable records. Biased estimates can occur when unusable units have significantly different characteristics from the usable ones. As in Wave 1, studies were completed to understand the non-response mechanism. Results showed that non-response units and unresolved units displayed different patterns and different rates were obtained for different characteristics of immigrants.

After numerous studies of the different rates and characteristics, it was fair to assume non-random response and resolved patterns. Both responding and non-responding units as well as resolved and unresolved units showed different patterns. Every non-random pattern must be corrected with the use of appropriate weight adjustment classes, taking into account the characteristics that lead to these different patterns. For example, if sex is an explanatory variable in the response prediction model, (i.e. different response rates for male and female), then sex must be used in the correction.

For these reasons, the adjustment weights were calculated in distinct steps for the responding units and for the resolved units as described in Section 12.3. Response and resolution models were used to construct the proper adjustment weights to correct for the fact that there were different response rates and different resolved rates. It also stresses the importance of using the final weights in any tabulation or analysis using the LSIC data. Any estimation done without the use of weights will produce biased results.

13.4 Coverage

Coverage is an indication of how a survey frame covers the target population or in the case of the LSIC, the population of interest. There could be over-coverage if the survey frame contains units that should not have been included, such as death, duplicates, or incorrect date of birth captured on the file. There could also be under-coverage, if the survey frame missed some units that should have been included. At Wave 1, there was a slight over-coverage which was corrected using a post-stratification technique on a more up-to-date file. In the absence of a more reliable source, the same file was used for Wave 2 (see Section 12.3.3). Thus, the size of the population of interest at Wave 2 is itself an estimate based on the Wave 1 data and the Wave 2 collection results.

14.0 Guidelines for Tabulation, Analysis and Release

This chapter of the documentation outlines the guidelines to be adhered to by users tabulating, analyzing, publishing or otherwise releasing any data derived from the survey microdata files. With the aid of these guidelines, users of microdata should be able to produce the same figures as those produced by Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

14.1 Rounding Guidelines

First, the distinction between rounding for reasons of protecting respondent confidentiality and rounding for the purpose of implied precision must be made. Rounding is often used as disclosure control, to prevent the linking of published results to individual respondents on a public use microdata file (PUMF). Because no PUMF will be produced for the Longitudinal Survey of Immigrants to Canada (LSIC), the linking of results is not a concern. However, the LSIC does release detailed geography, and, given the visible nature of LSIC respondents, weighted counts based upon sub-provincial geography must be rounded to the nearest multiple of fifty. Users of the LSIC microdata files must adhere to the following guidelines regarding the rounding of such estimates:

a) Estimates in the main body of a statistical table are to be rounded to the nearest multiple of fifty using the normal rounding technique.

b) Marginal sub-totals and totals in statistical tables are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 50 units using normal rounding. It is also acceptable, from a confidentiality point of view, to calculate the marginal sub-totals and totals using the rounded counts.

c) Averages, proportions, rates and percentages are to be computed from rounded components (i.e. numerators and/or denominators).

d) Sums and differences of aggregates are to be derived from their corresponding rounded components and then are to be rounded themselves to the nearest 50 units using normal rounding.

Rounding is also used so as not to imply greater precision than actually exists. In much of the published LSIC research produced by Statistics Canada, rounding, as described above, to the nearest hundred units is used. To ensure comparability between published results, users are urged to adhere to this practice. However, in instances where estimates to be published or otherwise released differ from corresponding estimates published by Statistics Canada, users should note the reason for such differences in the publication or release document(s).

14.2 Sample Weighting Guidelines for Tabulation

The sample design used for the LSIC was self-weighting. When producing simple estimates including the production of ordinary statistical tables, users must apply the final weight. If final weights are not used, the estimates derived from the microdata files cannot be considered to be representative of the survey population, and will not correspond to those produced by Statistics Canada. The weight assigned to each immigrant reflects the number of immigrants represented by a particular respondent.

Users should also note that some software packages may not allow the generation of estimates that exactly match those available from Statistics Canada, because of their treatment of the weight field (e.g. truncation or rounding of non-integer weights).

The LSIC Wave 2 file has been set up so that the longitudinal respondent is the unit of analysis. The weight that can be found on each record (WT2L) is an "immigrant" (the longitudinal respondent) weight. Analysis using the respondent's children, spouse, family or household as the unit of analysis cannot be carried out using the LSIC data. All research questions must be framed in terms of the longitudinal respondent.

14.3 Definitions of Types of Estimates: Categorical and Quantitative

Categorical Estimates

Categorical estimates are estimates of the number, or percentage of the surveyed population possessing certain characteristics or falling into some defined category. The number or the proportion of immigrants who plan to buy a house or an apartment in the next few years are examples of such estimates. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Examples of Categorical Questions:

Q: Do you or you and your family have plans to buy a house or an apartment in the next few years?
R: Yes / No / Not sure

Q: How many rooms are there where you live? (Include kitchen, bedrooms, finished rooms in the attic or basement, etc.) Do not count bathrooms, halls, vestibules and rooms used solely for business purposes.
R: One / Two / Three / Four / Five or more

Quantitative Estimates

Quantitative estimates are estimates of totals or of means, medians and other measures of central tendency of quantities based upon some or all of the members of the surveyed population. They also specifically involve estimates of the form $\hat{X}/\hat{Y}$ where $\hat{X}$ is an estimate of surveyed population quantity total and $\hat{Y}$ is an estimate of the number of persons in the surveyed population contributing to that total quantity.

An example of a quantitative estimate is the average monthly amount paid in housing costs. The numerator is an estimate of the total amount paid each month for the immigrants who live in dwelling units and the denominator is the number of immigrants who live in dwelling units.

Examples of Quantitative Questions

Q: How much do you or you and your family pay each month towards housing? (Include rent, taxes, heat, water, electricity, parking, condominium fees/mortgage, etc., but exclude telephone and cable.)
R: |_|_|_|_| $/month

Q: In this job, what is/was your wage or salary before taxes or other deductions?
R: |_|_|_|_|_|_| $

14.3.1 Tabulation of Categorical Estimates

Estimates of the number of immigrants with a certain characteristic can be obtained from the microdata file by summing the final weights of all records possessing the characteristic(s) of interest. These estimates may be cross-sectional or longitudinal. Proportions and ratios of the form $\hat{X}/\hat{Y}$ are obtained by:

a) summing the final weights of records having the characteristic of interest for the numerator $\left(\hat{X}\right)$,
b) summing the final weights of records having the characteristic of interest for the denominator $\left(\hat{Y}\right)$, then
c) divide estimate a) by estimate b) $\left(\hat{X} / \hat{Y}\right)$

14.3.2 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata file by multiplying the value of the variable of interest by the final weight for each record, then summing this quantity over all records of interest. For example, to obtain an estimate of the total amount paid monthly in housing costs, multiply the monthly amount of the immigrant's housing costs by the final weight for the record, then sum this value over all records for immigrants who live in dwelling units.

To obtain a weighted average of the form $\hat{X}/\hat{Y}$, the numerator ($\hat{X}$) is calculated as for a quantitative estimate and the denominator ($\hat{Y}$) is calculated as for a categorical estimate. For example, to estimate the average monthly amount paid for housing by immigrants living in dwelling units,

a) estimate the total monthly amount paid in housing costs $\left(\hat{X}\right)$ as described above,
b) estimate the number of immigrants who live in dwelling units $\left(\hat{Y}\right)$ by summing the final weights of all records for this category, then
c) divide estimate a) by estimate b) $\left(\hat{X} / \hat{Y}\right)$.

14.4 Guidelines for Statistical Analysis

The LSIC is based upon a complex sample design, with stratification, multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used. In order for survey estimates and analyses to be free from bias, the survey weights must be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures differs from that which is appropriate in a sample survey framework, with the result that, while in many cases the estimates produced by the packages are correct, the variance estimates that are calculated are poor. Approximate variances for simple estimates such as totals, proportions and ratios (for qualitative variables and for common domains) can be derived using the Wave 2 LSIC Coefficients of Variation Extraction Module (CVEM), which is provided as a companion tool. The CVEM is discussed in Section 15.3.

For other analysis techniques (for example, linear regression, logistic regression and analysis of variance), a method exists which can make the variances calculated by the standard packages more meaningful, by incorporating the unequal probabilities of selection. The method rescales the weights so that there is an average weight of 1. Commonly used analysis software (SAS and SPSS for example) often include options in many of their procedures that enable the rescaling of weights. However, variances calculated in this way do not account for the gains or losses in efficiency due to the stratification and clustering of the sample's design. Methods and software that allow for the appropriate estimation of variances are discussed Chapter 15.0.

14.5 Coefficient of Variation Release Guidelines

Before releasing and/or publishing any estimate from the LSIC, users should first determine the quality level of the estimate. The quality levels are *acceptable*, *marginal* and *unacceptable*. As discussed in Chapter 13.0, sampling and non-sampling errors both influence data quality. For the purposes of this document, however, estimate quality is based solely on the sampling error illustrated by the coefficient of variation, as shown in the table below.

First, the number of immigrants who contribute to the calculation of the estimate should be determined. If this number is less than 10, the weighted estimate cannot be released.

For weighted estimates based on sample sizes of 10 immigrants or more, users should determine the coefficient of variation of the estimate and follow the guidelines below. These quality level guidelines should be applied to weighted rounded estimates.

Quality Level Guidelines

Quality Level of Estimate	Guidelines
1) Acceptable	Estimates have: a sample size of 10 or more, and low coefficients of variation in the range of 0.0% to 16.5% No warning is required.
2) Marginal	Estimates have: a sample size of 10 or more, and high coefficients of variation in the range of 16.6% to 33.3%. Estimates should be flagged with the letter M (or some similar identifier). They should be accompanied by a warning to caution users about the high levels of error, associated with the estimates.
3) Unacceptable	Estimates have: a sample size of 10 or more, and very high coefficients of variation in excess of 33.3%. Statistics Canada recommends not to release estimates of unacceptable quality. However, if the user chooses to do so then estimates should be flagged with the letter U (or some similar identifier) and the following warning should accompany the estimates: "Please be warned that these estimates [flagged with the letter U] do not meet Statistics Canada's quality standards. Conclusions based on these data will be unreliable, and most likely invalid."

15.0 Variance Calculation

The Longitudinal Survey of Immigrants to Canada (LSIC) is a probabilistic survey, i.e. a sample has been selected to represent the target population. A given variability is inherent in any random selection. This variability is known as the sampling error, as described in Section 13.1. In addition, adjustments have been made to take into account non-responding and unresolved units which are part of the variability of the estimates. This chapter explains why it is important to calculate the variance and presents different tools to do so.

15.1 Importance of the Variance

The variance of an estimate is a good indicator of the quality of the estimate. A high variance estimate is considered unreliable. In order to quantify large variance, a relative measure of the variability is used, namely the coefficient of variation (CV). The coefficient of variation is defined as the ratio of the square root of the variance over the estimate. The square root of the variance is also known as a standard deviation. The coefficient of variation, as opposed to the variance, allows the analyst to compare estimates of different magnitudes along the same scale. As a result, it is possible to assess the quality of any estimate with the CV.

Most importantly variance or the CV is required for statistical tests such as hypothesis tests, which determine if two estimates are statistically different. Consequently, variance or CV calculation is mandatory.

Method to Obtain the Variance of an Estimate
It is almost impossible to derive an exact formula to calculate the variance for the LSIC due to the complex sample design, weight adjustments and post-stratification. A very good way to approximate the true variance is to use a replication method, namely the bootstrap method. This method is known to correctly approximate the true value of the variance. A file containing 1,000 bootstrap weights is available. Variance calculation using 1,000 bootstrap weights involves calculating the estimates with each of these 1,000 weights and then, calculating the variance of these 1,000 estimates.

Two user-friendly tools, both using the bootstrap weights, have been developed to help users calculate the variance and the CVs for their estimates. These tools are:

- **Macros to calculate the variance** using bootstrap weights, programmed for SAS and STATA users.
- **An Excel based CV extraction module** (CVEM) for totals and proportions, which produces approximate CVs for a large number of domains.

Like the LSIC macros, Bootvar is a program composed of macros which allow the computation of variance estimates using the bootstrap method. The program is generic, i.e., it is possible to use with data from any Statistic Canada's survey that releases bootstrap weights. The Bootvar program is available in SAS and SPSS formats.

Also, there are commercial software (SUDAAN, WesVar) that can produce variance estimates using the bootstrap weights. The advantage to these software is that, in addition to producing bootstrap variance estimates for a wider range of statistics, they allow for design-based corrections to other useful statistics. Use of the bootstrap weights with these software is discussed in *Using bootstrap weights with WesVar and SUDAAN*[8].

The use of one or more of these tools depends on the type of analysis and the level of precision required.

[8] http://www.statcan.ca/english/freepub/12-002-XIE/2004002/pdf/phillips.pdf

15.2 SAS and STATA Macros to Calculate the Variance Using the Bootstrap Weights

SAS and STATA macros have been developed to calculate the variance using the bootstrap weights. Variance calculation using these macros is more time consuming than the other method presented (i.e. CVEM). The user must first become familiar with the macros before using them. However, these macros have been developed in such a way that they are easy to use.

Despite the time required to run these macros, it is strongly recommended to use this method to calculate the variance of any estimates to be published. This method provides a more precise and accurate measure of the variance.

15.3 Excel Based Coefficient of Variation Extraction Module

The second tool available for users to obtain approximate coefficients of variation is the Excel based CV extraction module (CVEM). This application, developed with Excel macros and accessed through a user-friendly interface, allows user to extract the desired information in two ways. One is by describing the domain of interest with the nine available variables, and the other is by specifying the size of the domain. The information displayed consists of the proportion estimate, the number of respondents in the specified domain, the estimated population in that domain, basic statistics and the coefficient of variation for the selected proportion. Here, a domain is defined as being the cross-tabulation of the variables listed in the table in Section 15.3.1.

Over 44,000 domains are covered by the set of spreadsheets, giving an approximate CV for eight different proportions in each of the domains, for a total of over 352,000 CV's. Simulations were run to calculate variances, coefficients of variation and confidence intervals at the 95% level for different proportions, i.e. 1%, 5%, 10%, 15%, 20%, 30%, 40% and 50%. These proportions were based on population distribution. For a given repetition, the observed proportion in the random sample can be different from that of the targeted proportion. Therefore the mean of 100 repetitions was used to account for that variability.

15.3.1 Statistics Canada Quality Standards

Users should note that for disclosure issues, when using a dichotomous variable, both the sample size and the CV should be publishable simultaneously. Users should always ensure the quality of the estimates, especially for smaller proportions obtained from small domains. To help users identify high CVs, color coding has been used in the Excel application when displaying a CV. Using the markers described below, the colors used are red for CV's in excess of 33.3% and yellow for the ones in the range of 16.6% to 33.3%. More details are provided in the CVEM User's Guide. Below is a list of the variables available in the CVEM.

Field	Description
Class of immigrant	
Age group	
Geographical residence	
Place of birth	
Gender	
Marital status	
Employment status	
Highest level of education	
Knowledge of official languages	
Target proportion	The theoretical proportion used to simulate a variable. Can take the values 1%, 5%, 10%, 15%, 20%, 30%, 40% or 50%
Y hat	The mean of 100 calculated proportions. This figure should be close to the target proportion.
N	The average sample size of the specified domain from 100 repetitions.
Bs_var	The mean of 100 variances for the specified domain.
Bs_sd	The mean of 100 standard errors for the specified domain.
Cil95	The mean of 100 at the 95% confidence interval lower boundary.
Ciu95	The mean of 100 at the 95% confidence interval upper boundary.

As a reference, the following quality standards should be used:

1) An estimate is said to be **acceptable** if it has a sample size of 10 or more and low coefficient of variation in the range of 0.0% to 16.5%.

2) An estimate is said to be **marginal** if it has a sample size of 10 or more and high coefficient of variation in the range of 16.6% to 33.3%. This estimate should be accompanied by a warning to caution subsequent users about the high level of error, associated with the estimate.

3) An estimate is said to be **unacceptable** if it has a sample size of 10 or more and very high coefficient of variation in excess of 33.3%. Statistics Canada recommends not to release estimates of unacceptable quality (see Section 14.5).

For more information see the publication *Statistics Canada Quality Guidelines,* Catalogue no. 12-539-XIE.

15.4 How to Derive the Coefficient of Variation for Categorical Estimates

Rule 1: Estimates of Number of Immigrants Possessing a Characteristic (Aggregates)

The coefficient of variation depends only on the size of the estimate itself. It is safe to say that an estimate's CV is close (though slightly greater) than the proportion it represents. Hence, to get an approximation of an estimate's CV, users could use the CVEM by specifying the domain's size and deriving the appropriate proportion. For example, suppose we have an estimate *Y* hat = *30,000* individuals possessing a certain characteristic. If we are to compare them to the *100,000* people in the domain of interest, then the CV for *Y* hat should be close to the proportion i.e. *30,000 / 100,000 = 30.0%.* To have an a more precise CV, the programs that use the bootstrap weights should be used. Bootstrap programs are available for SAS and STATA users.

Rule 2: Estimates of Proportions or Percentages of Immigrants Possessing a Characteristic

The CV's calculated in the CVEM are for proportions. Hence, they can be used directly as they are given on the spreadsheet.

Rule 3: Estimates of Differences Between Aggregates, Percentages and Ratios

To obtain the CV for a difference, the Bootstrap programs are best suited as there is no easy way to derive it from each of the individual CV's. The programs offer the possibility to derive CV's for differences of totals and ratios.

Rule 4: Estimates of Ratios

If the denominator of a ratio is considered as a "domain size", one can use the CVEM just as it is used in Rule 2. Otherwise, the Bootstrap programs can be used by defining properly the numerator and the denominator.

15.5 How to Use the Coefficient of Variation to Obtain Confidence Limits

Although coefficients of variation are widely used, a more intuitively meaningful measure of sampling error is the confidence interval of an estimate. A confidence interval constitutes a statement on the level of confidence that the true value for the population lies within a specified range of values. For example, a 95% confidence interval can be described as follows:

> If sampling of the population is repeated indefinitely, with each sample leading to a new confidence interval for an estimate, then in 95% of the samples, the interval will cover the true population value.
>
> Using the standard error of an estimate, confidence intervals for estimates may be obtained under the assumption that under repeated sampling of the population, the various estimates obtained for a population characteristic are normally distributed about the true population value. Under this assumption, the chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and about 99 out of 100 that the differences would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

Confidence intervals for an estimate, $\hat{X}$, are generally expressed as two numbers, one below the estimate and one above the estimate, as $\left(\hat{X} - k, \hat{X} + k\right)$ where k is determined depending upon the level of confidence desired and the sampling error of the estimate.

The 95% confidence intervals for an estimate are available directly in the CV spreadsheet. If the user wants to determine other confidence intervals, the following formula will convert to a confidence interval $\left(CI_{\hat{x}}\right)$:

$$CI_{\hat{x}} = \left(\hat{X} - t\hat{X}\alpha_{\hat{x}}, \hat{X} + t\hat{X}\alpha_{\hat{x}}\right)$$

where $\alpha_{\hat{X}}$ is the determined coefficient of variation for $\hat{X}$ and

t = 1 if a 68% confidence interval is desired;
t = 1.6 if a 90% confidence interval is desired;
t = 2.6 if a 99% confidence interval is desired.

Warning Note on Confidence Intervals
Release guidelines which apply to the estimate also apply to the confidence interval. For example, if the estimate is "marginal", then the confidence interval is marginal and should be accompanied by a warning note to caution subsequent users about high levels of error, associated with the estimate.

Example of Using the Coefficient of Variation to Obtain Confidence Limits
A 90% confidence interval for the estimated proportion of women having a university degree would be calculated as follows:

$\hat{X}$ = 47.4% (or expressed as a proportion 0.474)

t = 1.6

$\alpha_{\hat{x}}$ = 1.21% (0.0121 expressed as a proportion) is the coefficient of variation of this estimate as derived using the bootstrap weights.

$CI_{\hat{x}}$ = {0.474 - (1.6) (0.474) (0.0121), 0.474 + (1.6) (0.474) (0.0121)}

$CI_{\hat{x}}$ = {0.474 – 0.009, 0.474 + 0.009}

$CI_{\hat{x}}$ = {0.465, 0.483}

Hence, with a 90% level of confidence, it can be said that between 46.5% and 48.3% of women have a university degree.

15.6 Hypothesis Testing (t-test)

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The sample estimates can be numbers, averages, percentages, ratios, etc. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

Let $\hat{X}_1$ and $\hat{X}_2$ be sample estimates for two characteristics of interest. The standard error for the difference $\hat{X}_1 - \hat{X}_2$ can be obtained through the programs that use the bootstrap weights. Let the standard error on the difference be $\sigma_{\hat{d}}$.

If $t = \dfrac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}}$ is between -2 and 2, then no conclusion about the difference between the characteristics is justified at the 5% level of significance. If however, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 0.05 level. That is to say that the difference between the estimates is significant.

15.7 Coefficients of Variations for Quantitative Estimates

For quantitative estimates, special tables would have to be produced to determine their sampling error. Since most of the variables for the LSIC are primarily categorical in nature, this has not been done.

As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding category estimate (i.e., the estimate of the number of persons contributing to the quantitative estimate). If the corresponding category estimate is not releasable, the quantitative estimate will not be either. For example, the coefficient of variation of the total number of hours of class for women attending university courses would be greater than the coefficient of variation of the corresponding proportion of women attending university courses. Hence if the coefficient of variation of the proportion is not releasable, then the coefficient of variation of the corresponding quantitative estimate will also not be releasable.

Pseudo Replication
Coefficients of variation of such estimates can be derived as required for a specific estimate using a technique known as pseudo replication. This involves dividing the records on the microdata files into subgroups (or replicates) and determining the variation in the estimate from replicate to replicate. Users wishing to derive coefficients of variation for quantitative estimates may contact Statistics Canada for advice on the allocation of records to appropriate replicates and the formulae to be used in these calculations.

15.8 Approximate Quality Release Cut-offs

The tables below provide the approximate release cut-offs for two selected domains. These population estimates provide a rough indication of acceptable, marginal and unacceptable domain sizes. They are meant to be used as approximate guidelines only. Users are still responsible to calculate precise CVs before releasing results. The use of the CVEM is strongly recommended for better precision.

Approximate Release Cut-offs by Class of Immigrant

Class of Immigrants	Acceptable CV 0.0% to 16.5%	Marginal CV 16.6% to 33.3%	Unacceptable CV > 33.3%
Family	800 & over	200 to < 800	under 200
Economic	550 & over	160 to < 550	under 160
Refugees	260 & over	75 to < 260	under 75
Total	**520 & over**	**140 to < 520**	**under 140**

Approximate Release Cut-offs by Geographical Regions

Province	Acceptable CV 0.0% to 16.5%	Marginal CV 16.6% to 33.3%	Unacceptable CV > 33.3%
Quebec	550 & over	150 to < 550	under 150
Ontario	610 & over	150 to < 610	under 150
Alberta	380 & over	90 to < 380	under 90
British Columbia	480 & over	170 to < 480	under 170
Other	370 & over	190 to < 370	under 190
Canada	**510 & over**	**140 to < 510**	**under 140**

16.0 Record Layout with Univariate Frequencies

***Available in the Research Data Centres only.**