Catalogue no. 75F0002MIE — No. 007
ISSN 1707-2840
ISBN 978-0-662-47062-5

**Research Paper**

## Income Research Paper Series

# Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics

2003

by Sylvie LaRoche

Households Survey Methods Division
Jean Talon Building, Ottawa, K1A 0T6

Telephone: 1-613-951-7355

Statistics Canada Statistique Canada

Canada

## How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 613-951-7355; 888-297-7355; income@statcan.ca).

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.ca or contact us by e-mail at infostats@statcan.ca or by phone from 8:30am to 4:30pm Monday to Friday at:

| | |
|---|---|
| **Toll-free telephone** (Canada and the United States): | |
| Inquiries line | 1-800-263-1136 |
| National telecommunications device for the hearing impaired | 1-800-363-7629 |
| Fax line | 1-877-287-4369 |
| Depository Services Program inquiries line | 1-800-635-7943 |
| Depository Services Program fax line | 1-800-565-7757 |
| | |
| **Statistics Canada national contact centre:** | 1-613-951-8116 |
| Fax line | 1-613-951-0581 |

## Information to access the product

This product, catalogue no. 75F0002MIE, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1-800-263-1136. The service standards are also published on www.statcan.ca under About us > Providing services to Canadians.

Statistics Canada

Income Statistics Division

**Income Research Paper Series**

# Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics

2003

Published by authority of the Minister responsible for Statistics Canada



October 2007

Catalogue no. 75F0002MIE, no. 007
ISSN 1707-2840
ISBN 978-0-662-47062-5

Frequency: occasional

Ottawa

La version française de cette publication est disponible sur demande (nº 75F0002MIF au catalogue).

---

**Note of appreciation**

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

**Abstract**

The Survey of Labour and Income Dynamics (SLID), introduced in the 1993 reference year, is a longitudinal panel survey of individuals. The purpose of the survey is to measure changes in the economic well-being of individuals and the factors that influence these changes. SLID's sample is divided into two overlapping panels, each six years in length. Longitudinal surveys like SLID are complex due to the dynamic nature of the sample, which in turn is due to the ever-changing composition of households and families over the years. For each reference year, SLID produces two sets of weights: one is representative of the initial population (the longitudinal weights), while the other is representative of the current population (the cross-sectional weights). Since 2002, SLID has been producing a third set of weights which combines two panels that overlap to form a new longitudinal sample. The new weights are referred to as combined longitudinal weights.

For the production of the cross-sectional weights, SLID combines two independent samples and assigns a probability of selection to individuals who joined the sample after the panel was selected. Like cross-sectional weights, longitudinal weights are adjusted for non-response and influential values. In addition, the sample is adjusted to make it representative of the target population. The purpose of this document is to describe SLID's methodology for the longitudinal and cross-sectional weights, as well as to present problems that have been encountered, and solutions that have been proposed. For the purpose of illustration, results for the 2003 reference year are used. The methodology used to produce the combined longitudinal weights will not be presented in this document as there is a complete description in Naud (2004).

## Table of contents

# 1 Preface

The original version of this document, which describes in great detail the longitudinal and cross-sectional weights of the Survey of Labour and Income Dynamics (SLID), was written by Isabelle Lévesque and Sarah Franklin, and was published in 2000. Since then, several changes and new developments were introduced into the methodology for both longitudinal and cross-sectional weighting. Therefore, it was necessary to update this document. Here is a summary of the major changes that were made to the weighting methodology, which are described in this new version:

- In terms of non-response adjustments, an adjustment factor has now been attributed to the weighting of children and out-of-scope persons (see sections 4.2 and 4.3)

- In cross-sectional weighting, the adjustment for interprovincial migration is now performed after the non-response adjustment. Before, this step was performed following the weight share (see section 5.3)

- Modifications were made to the methodology of weight share for labour weights. This step is now performed in a way that better-preserves the non-response adjustment, and thus the relationships between the longitudinal individuals in a household (see section 5.5.2)

- When looking at overlapping panels during cross-sectional weighting, it is necessary to identify immigrants and to know their date of entry, in order to determine whether they could have been selected in one of the overlapping panels. However, the immigration questions are asked only of persons 16 or older. A strategy for the imputation of immigration status and entry date for children under 16 was put in place in order to better assign a panel allocation factor (see section 5.4)

- Changes were also made to the way in which panel allocation factors are applied. The factor $p$ is now applied only to longitudinal individuals of the oldest panel who were eligible during the selection of the most recent panel (see section 5.4)

- The post-stratification step was replaced by auxiliary data calibration following the addition of new control totals. For longitudinal weighting, a two-step calibration is now being performed which permits the use of control totals similar to the ones used in cross-sectional weighting (see sections 4.5 and 5.7 for the longitudinal and cross-sectional weighting calibration, respectively)

## 2 Introduction

The Survey of Labour and Income Dynamics (SLID), introduced in the 1993 reference year, is a longitudinal panel survey of individuals. Its goal is to measure changes in the economic well-being of individuals and the factors that influence those changes, especially key factors in the areas of demographic and family characteristics and labour market activity. Its end product is a set of microdata files. For an overview of the survey, see Lavigne and Michaud (1998).

Originally, SLID was intended to provide longitudinal data to be used in producing longitudinal estimates and analyses. Its cross-sectional dimension has now become just as important as its longitudinal one, in part because the Survey of Consumer Finances (SCF), a cross-sectional survey, has been merged with SLID (Cotton et al., 1999). The two surveys were integrated for the 1998 reference year. For more information about SCF, see Statistics Canada (1997).

Weighting a longitudinal survey can sometimes be a major methodological challenge, not only because of its longitudinal and cross-sectional dimensions, but also because of the dynamic nature of the panel (which is in turn due to the ever-changing composition of families and households). Many weighting problems have arisen over the years: the longitudinal combination of overlapping panels, the introduction of new calibration groups, updates in weighting programs to make them more versatile and effective, on-going research to improve our methodology, and so on. All these factors add to the complexity of SLID's longitudinal and cross-sectional weighting process.

The main purpose of this report is to describe the various steps in the longitudinal and cross-sectional weighting process for the 2003 reference year. Our second objective is to inform the reader of future developments and forthcoming changes affecting weighting. The report consists of five sections. The survey's methodology is discussed in section 3. Section 4 presents the steps in the longitudinal weighting process, from determination of the initial weights to calculation of the final ones, while section 5 covers the various steps in the cross-sectional weighting process. It should be noted that the longitudinal and cross-sectional aspects are combined in the description of some of the weighting steps. Section 6 looks to the future, and section 7 contains the report's conclusion.

This report is an update of Lévesque and Franklin's 2000 report on SLID's weighting. An updated version was deemed necessary because the longitudinal and cross-sectional weighting methodology has undergone a number of changes and innovations since 2000.

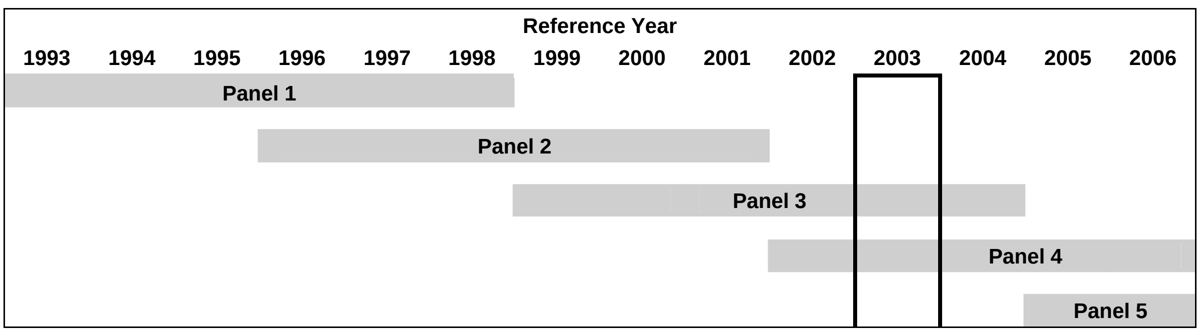
# 3 Survey methodology

The SLID sample is composed of two panels, each six years (or six waves) in length. Each panel covers the population living in the ten Canadian provinces at a specific date. The population excludes persons living on Indian reserves, full-time members of the Canadian Armed Forces, and persons who had been living in an institution for more than six months. All references to the population of the ten Canadian provinces in this document automatically imply the above exclusions.

Panel 1, selected in January 1993, covers the population living in the ten Canadian provinces on December 31, 1992. Panel 2 was introduced in 1996, and since then a new panel has been selected every three years to replace the oldest panel. The purpose of this rotation is to make the cross-sectional sample more representative (to reflect the new households added to the population in the preceding three years), to mitigate the effect of attrition, and to reduce response burden. Hence, there is always a three-year overlap between successive panels.

The period used in this report to illustrate the weighting strategy – the 2003 reference year – is covered by Panels 3 and 4, which represent the population of Canada's ten provinces on December 31 of 1998 and 2001 respectively. Figure 3.1 provides an overview of panel selection and overlap.

**Figure 3.1 Overlap of SLID panels**



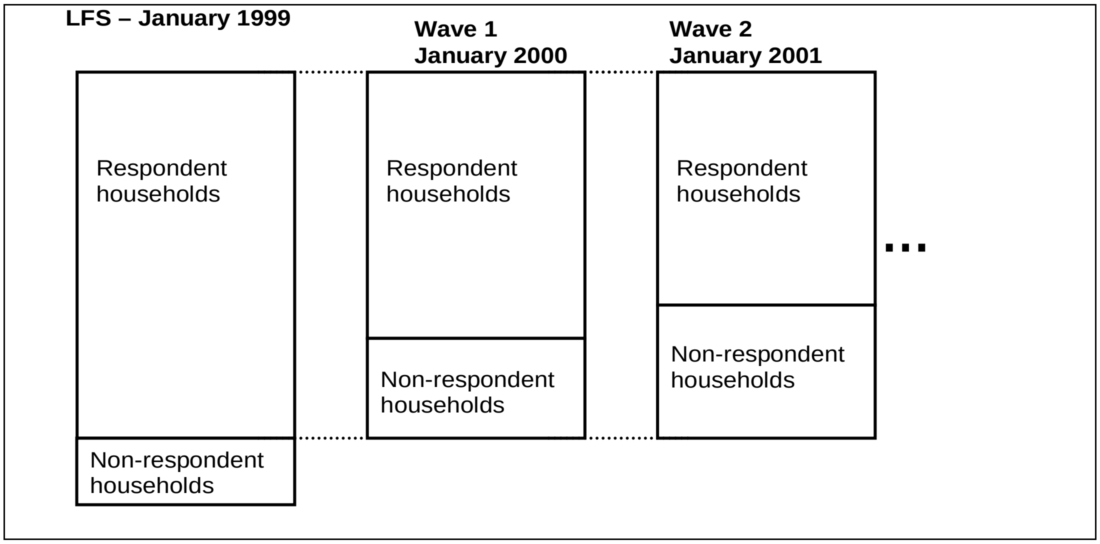

Each SLID panel consists of a subsample of about 17,000 households (roughly 40,000 people) from the Labour Force Survey (LFS). The LFS sample is drawn from an area frame in accordance with a multistage sampling design. The LFS has six panels (rotation groups); each panel remains in the sample for six months, and one panel is replaced each month. The last-stage sampling unit is the dwelling. Thus, all members of the households occupying the selected dwellings are included in the LFS sample. For further information about the LFS, see the reports on its methodology (Singh *et al.*, 1990; Gambino *et al.*, 1998).

The households selected for SLID belong to the panels rotated out of the LFS sample in the first two months of SLID's first reference period (January and February 1999 for Panel 3, and January and February 2002 for Panel 4). Of all the households that make up the outgoing LFS rotation groups in January or February, SLID selects only those which responded in the final month of their LFS participation. The final LFS interview serves as the introductory contact for SLID, and as an opportunity to collect contact information to facilitate tracing in wave 1 data collection.

In wave 1, a preliminary interview is conducted with each person selected by SLID. Basic information about work experience, level of education, and family and personal background is collected during the interview. For Panel 1, the preliminary SLID interview was conducted at the same time as the final LFS interview. For more details on the procedures followed when Panel 1 was introduced, see Lévesque and Franklin (2000). Figure

3.2 illustrates SLID's household selection process for Panel 3. More details on the selection of SLID panels are provided in subsection 4.1.

**Figure 3.2 Selection of households for SLID longitudinal sample for panel 3**




Each person selected for SLID is then interviewed twice a year for six years. Each year, information is collected about family and household composition, labour market activity, education, housing and the previous year's income. The labour market, education and housing questions are asked in January, and the income questions in May. Income data are collected in May because by then, most people have filed their income tax returns, and their income information is likely to be more accurate and complete. To reduce the response burden, SLID allows respondents to avoid the May interview by preauthorizing Statistics Canada to access their Canada Revenue Agency tax file. In the 2005 collection process (for the 2004 reference year), the May interview was combined with the January interview to reduce collection costs. The impact that this change has had on the quality of income data and the tax data permission rates is still to be assessed.

When a panel is selected for SLID, every member of the chosen households, regardless of age, becomes part of the panel's longitudinal sample. These persons are considered part of the longitudinal sample for the panel's full six years, even if they move, die, are institutionalized or become full-time members of the Armed Forces living in military quarters. No one else can join the longitudinal sample for this panel. Thus, for each panel, the longitudinal sample is formed when the panel is selected and remains unchanged for the life of the panel. The sample's target population will stay the same throughout the six years that the panel is part of the survey.

For longitudinal purposes, the unit of interest is the individual; because of its constantly changing nature, a household can hardly be used as a tool for longitudinal analysis. For cross-sectional purposes, we are interested in both the individual and the household. Since SLID is concerned with the characteristics not only of longitudinal persons but also of households, everyone who lives with at least one longitudinal individual is interviewed. Hence, the cross-sectional sample for a given year is composed of all longitudinal persons in scope on December 31 of the corresponding reference year, and all persons living with them on that date. An individual is considered in-scope on December 31 of a given reference year (cross-sectionally in scope) if on that date, he or she lives in one of the ten Canadian provinces, does not live on an Indian reserve, has not been living in an institution for more than six months, and is not a full-time member of the Canadian Armed Forces living in military barracks. Every cross-sectionally in-scope longitudinal person will also be part of the cross-sectional sample. Interviewees who are not part of the longitudinal sample are called cohabitants. Adding cohabitants to the cross-sectional sample helps make it more representative (Latouche and Michaud, 1997).

To meet the above-mentioned requirements, longitudinal and cross-sectional weights are produced each year. Longitudinal weighting is designed to generate estimates that are representative of the population of Canada's ten provinces at the time the longitudinal sample is selected, while cross-sectional weighting produces estimates that are representative of the provinces' population on December 31 of a given reference year. For example, the estimates based on Panel 3's longitudinal sample are representative of the provinces' population on December 31, 1998, and the estimates produced with Panel 4's longitudinal sample relate to the provinces' population on December 31, 2001. The estimates based on the cross-sectional sample for the 2003 reference year represent the provinces' population on December 31, 2003.

In January 1999, 43,683 persons were selected for Panel 3's longitudinal sample. In January 2002, 42,232 persons were chosen for Panel 4's longitudinal sample. The cross-sectional sample for the 2003 reference year consists of 75,206 individuals, 65,329 of whom are longitudinal persons.

Table 3.1 provides an overview of how the longitudinal samples have changed over time for Panels 3 and 4. Table 3.2 shows the composition of the cross-sectional samples of all reference years considered (1999 to 2003 inclusive).

**Table 3.1 Composition of the longitudinal samples of panels 3 and 4 for the 1999 to 2003 reference years**

| | Reference year | | | | |
|---|---|---|---|---|---|
| | 1999 | 2000 | 2001 | 2002 | 2003 |
| | Panel 3 - Persons | | | | |
| **Respondents[1]** | number | | | | |
| In-scope on December 31 of a given reference year[2] | 36,158 | 35,340 | 34,892 | 32,922 | 31,214 |
| Living outside the ten Canadian provinces | 130 | 234 | 336 | 400 | 377 |
| Institutionalized | 97 | 169 | 255 | 437 | 514 |
| Deceased | 274 | 496 | 761 | 1,010 | 1,283 |
| Duplicate, error | 0 | 0 | 0 | 3 | 9 |
| **Non-respondents[3]** | | | | | |
| Hard refusal, persons in households not traced | 7,024 | 7,444 | 7,439 | 8,911 | 10,286 |
| **Total** | 43,683 | 43,683 | 43,683 | 43,683 | 43,683 |
| | Panel 4 - Persons | | | | |
| **Respondents[1]** | number | | | | |
| In-scope on December 31 of a given reference year[2] | … | … | … | 33,792 | 34,115 |
| Living outside the ten Canadian provinces | … | … | … | 108 | 241 |
| Institutionalized | … | … | … | 141 | 265 |
| Deceased | … | … | … | 255 | 515 |
| Duplicate, error | … | … | … | 0 | 14 |
| **Non-respondents[3]** | | | | | |
| Hard refusal, persons in households not traced | … | … | … | 7,936 | 7,082 |
| **Total** | … | … | … | 42,232 | 42,232 |

0 true zero or a value rounded to zero
... not applicable
1. Respondent longitudinal individuals will have a positive longitudinal weight.
2. Still living in the ten Canadian provinces, not deceased, not living in an institution for more than the last six months.
3. Non-respondent longitudinal individuals will have a longitudinal weight equal to 0. This category includes hard refusals, cases not contacted during collection and cases not returned to collection because they were non-respondent for two consecutive years.

### Table 3.2 Composition of cross-sectional samples, by panel for the 1999 to 2003 reference years

| | Reference year | | | | |
|---|---|---|---|---|---|
| | 1999 | 2000 | 2001 | 2002 | 2003 |
| | | | Panel 2 - Persons | | |
| **Respondents** | | | number | | |
| Longitudinal persons | 34,426 | 32,224 | 31,451 | … | … |
| Cohabitants | 5,673 | 6,210 | 7,306 | … | … |
| Subtotal | 40,099 | 38,434 | 38,757 | … | … |
| | | | Panel 3 - Persons | | |
| **Respondents** | | | number | | |
| Longitudinal persons | 36,158 | 35,340 | 34,892 | 32,922 | 31,214 |
| Cohabitants | 1869 | 3,072 | 4,830 | 5,753 | 6,406 |
| Subtotal | 38,027 | 38,412 | 39,722 | 38,675 | 37,620 |
| | | | Panel 4 - Persons | | |
| **Respondents** | | | number | | |
| Longitudinal persons | … | … | … | 3,3792 | 34,115 |
| Cohabitants | … | … | … | 2,012 | 3,471 |
| Subtotal | … | … | … | 35,804 | 37,586 |
| **Total - Cross-sectional respondents**[1] | 78,126 | 76,846 | 78,479 | 74,479 | 75,206 |

… not applicable
1. These individuals will have a non-zero cross-sectional weight for the corresponding reference year.

The non-respondent group has little effect on the composition of the cross-sectional samples (Table 3.2), since the non-response adjustment is based on the longitudinal sample (for more detailed information on the non-response adjustment for the longitudinal and cross-sectional weights, see sections 4.3 and 5.2). In addition, it is difficult to obtain an accurate picture of non-respondent cohabitants, since it is often impossible to determine the current composition of a non-respondent household. Cohabitants can leave the household or join it, and we have no way of knowing about it.

SLID's mandate is to collect longitudinal and cross-sectional data on labour-market and income characteristics, and to make the data available to users inside and outside Statistics Canada. To that end, SLID produces two sets of microdata files: one anonymized to preserve confidentiality (external files), and one not anonymized (internal files).

Once the data for a given reference year have been processed, Statistics Canada users can access the relational database to perform their cross-sectional and longitudinal analyses. All information in the database is confidential since it has not been anonymized. A data retrieval system has also been developed to make it easier for Statistics Canada users to access the survey data (\\lhs2\clientserv\Income\SLID\SLIDCopy\SLIDRet_Installation).

External users have access to the external microdata files, which are usually produced for each reference year. Releasing such files takes more work because they must be treated to preserve confidentiality. In particular, all disclosure control procedures are reviewed and improved in order to ensure data confidentiality. Longitudinal microdata files for external use were produced for the 1993 and 1994 reference periods (see Lévesque and Franklin, 2000). However, because of the difficulties involved in preserving data confidentiality in such files, only cross-sectional microdata files were released for subsequent periods. The dissemination strategy is now to produce a public-use, cross-sectional microdata file for each reference year. To reduce the risk of disclosure associated with such files, they are now composed of a subsample of households from the original cross-sectional sample. The data are reweighted to ensure that the subsample is representative of the target population for a given reference period.

Thus, the cross-sectional files (which relate to the reference year in question) contain a subsample of all persons (longitudinal persons and cohabitants) in respondent households that are in-scope on December 31 of the reference year concerned.

# 4 Longitudinal weighting

This section describes the methodology used to weight the longitudinal persons in each wave of SLID. For illustrative purposes, we will use examples from wave 5 of Panel 3 and wave 2 of Panel 4. It is important to keep in mind that the longitudinal weights are intended to represent the population of Canada's ten provinces on December 31, 1998, for Panel 3 and on December 31, 2001, for Panel 4. The two sets of longitudinal weights are produced separately for the two panels[1]. It should be noted that since 2002, a new group of longitudinal weights, combining two overlapping panels, was also produced. For more details on the methodology used to create this group of weights, please refer to Naud (2004)

The longitudinal weighting process consists of the five steps discussed in the following subsections: determination of initial weights (4.1), classification of longitudinal individuals (4.2), non-response modelling and adjustment (4.3), adjustment for influential values (4.4), and lastly calibration, to adjust some of the survey estimates to known population totals (4.5).

## 4.1 Determination of initial weights

For the third SLID sample (Panel 3), all households that were members of the LFS panels rotated out in January and February 1999, who responded to the LFS in January and February 1999 respectively, were selected (17,085 households, for a total of 43,683 persons). The preliminary interviews were conducted at the same time as the wave 1 interviews (i.e., in January 2000), as was the case when Panel 2 was introduced.[2]

The sample for Panel 4 was selected in the same way as the sample for Panel 3. That is, all households that were members of the LFS panels rotated out in January and February 2002 who responded to the LFS in January and February 2002, respectively, were selected (16,942 households, for a total of 42,232 persons).

The initial longitudinal weights for Panels 3 and 4 are equal to the households' inverse selection probability. Thus, all longitudinal members of a household (the one to which they belonged at the time the sample was selected) have the same initial weight.

Since the SLID sample is drawn from the LFS sample, the initial weight will be as follows:

$$w_{\text{initial}, p_i} = w_{\text{LFS}} * 3$$

where $w_{\text{initial}, p_i}$ = initial longitudinal weight for Panel *i, i=3,4*

$w_{\text{LFS}}$ = LFS weight adjusted for non-response (LFS subweight)

3 = inverse selection probability of the LFS rotation groups (2 out of 6 groups were selected)

Table 4.1 shows the median initial longitudinal weight for each province at the time the panels were selected, i.e., the median of the $w_{\text{initial}, p_i}$, for i=3, 4.

---

1. Previously, SLID produced two sets of longitudinal weights: a set of internal weights and a set of external weights. Since the external weights are no longer used, only the set of internal weights is produced now. For more details on how the external weights were generated, see Lévesque and Franklin (2000).
2. The procedures for selecting Panel 1 were different. See Lévesque and Franklin (2000) for more details on how the initial weights for that panel were determined.

**Table 4.1 Median initial longitudinal weights for panels 3 and 4**

| Province | Initial longitudinal Panel 3 | Initial longitudinal Panel 4 |
|---|---|---|
| | weight | |
| Newfoundland and Labrador | 237 | 238 |
| Prince Edward Island | 101 | 103 |
| Nova Scotia | 288 | 279 |
| New Brunswick | 279 | 265 |
| Quebec | 567 | 561 |
| Ontario | 560 | 566 |
| Manitoba | 336 | 329 |
| Saskatchewan | 280 | 282 |
| Alberta | 692 | 796[1] |
| British Columbia | 843 | 834 |

1. The larger difference for the Alberta province could be caused by the addition or removal of high population growth clusters from the LFS sample during that period.

## 4.2 Classification of longitudinal individuals

For each wave, the first step in longitudinal weighting is to place the longitudinal individuals in one of the following categories: nonrespondents (in respondent or nonrespondent households), cross-sectionally in-scope respondents, and cross-sectionally out-of-scope persons. A household is defined as respondent if at least one household member responded to at least one of the two interviews (the labour interview or income interview). A household is considered non-respondent if all household members are non-respondents at both interviews. It has also been decided that if information about a household's composition is missing for more than a year, the household will be deemed non-respondent. A person is deemed cross-sectionally out-of-scope if, on December 31 of the reference year, he or she was deceased, had been in an institution for more than six months, or was not living in one of Canada's ten provinces.

The classification of longitudinal individuals is necessary in order to determine which ones should be assigned a non-zero longitudinal weight. A longitudinal weight of zero is applied to individuals in non-respondent households. All other individuals initially selected for the longitudinal sample have a non-zero longitudinal weight, even if they are cross-sectionally out-of-scope (since they are considered to be respondents). It should be noted that for weighting purposes, SLID defines non-respondent persons in respondent households as respondent. This is to ensure weighting consistency within households (the data for non-respondent persons in respondent households are imputed).

The classification of longitudinal individuals also determines whether an individual's weight should be adjusted for non-response and identifies those individuals who will be used in the non-response adjustment model. The purpose of this model is to find the value by which respondents' weights must be adjusted to make them representative of all respondents and non-respondents. During the 2005 historical revision, changes were made in the non-response methodology. They were applied retroactively to all SLID reference years and affect primarily children in respondent households and out-of-scope persons. A brief overview of the changes is presented below. For more information about the previous non-response adjustment methodology, see Lévesque and Franklin (2000).

In the old methodology, children's and out-of-scope persons' weights were not adjusted for non-response. In the case of children's weight, the reasons were that there was too little information about children to properly model non-response, and that a non-response adjustment was made implicitly in post-stratification. However, slippage studies on SLID have shown that the absence of a non-response adjustment in children's weights was affecting household representation prior to calibration. Households with a larger number of children were much less well represented, which subsequently resulted in greater distortion of the weights at the calibration stage. Because so little information is available about children, and because they generally do not make the decisions regarding the household's participation in the survey, they are excluded from non-response modelling. Nevertheless, they receive a non-response adjustment equivalent to the average adjustment of longitudinal adults living in the household.

Furthermore, it was discovered that, contrary to assumptions, out-of-scope persons are not overrepresented. In fact, new studies found that the group was actually underrepresented (Godbout, 2004). Because of these findings and the fact that some non-respondents are probably out of scope, out-of-scope persons are now included in the non-response adjustment, though they are excluded from non-response modelling.

A non-response adjustment is usually applied to the weights of all respondent units. In SLID's case, a non-response adjustment factor is applied to the weights of all persons who are considered respondents.

Table 4.2 shows the various response categories of the longitudinal individuals in wave 5 of Panel 3 and wave 2 of Panel 4. The *Non-response modelling* column indicates which categories of individuals are used in the non-response adjustment model. The last column shows how the weights of individuals in the various categories will be processed.

## 4.3 Non-response modelling and adjustment

Having generated the initial weights, we now have to compute a non-response adjustment factor for every individual deemed to be a respondent, except for children in respondent households and cross-sectionally out-of-scope persons. As noted earlier, the two panels are processed separately. The non-response adjustment factor for an individual is defined as the inverse response-rate of the response homogeneity group (RHG) to which the individual belongs ($1/R_{RHG}$). RHGs are made by grouping together all of the individuals with similar non-response characteristics. If the RHGs are defined so that non-response is completely random within each group, then the non-response bias is negligible (Tambay et al., 1998). To find an appropriate adjustment value to make respondent weights representative of all respondents and non-respondents, we use non-response modelling.

The first step in non-response modelling is to determine which explanatory variables linked to non-response need to be in the model in order to correctly predict the dichotomous variable *response* (whose possible values are either *respondent* or *non-respondent*). There are two possible ways of carrying out the non-response adjustment. The adjustment can be based on the initial characteristics of the individuals in the sample or on the most recent available characteristics. Both methods have advantages and disadvantages.

**Table.4.2 Classification of longitudinal individuals in wave 5 of panel 3 and wave 2 of panel 4**

| Classification of individuals | Panel 3 wave 5 | Panel 4 wave 2 | Non-response modelling | $W_{adjust}$ (adjusted) |
|---|---|---|---|---|
| **Children aged 0 to 15(on Dec. 31, 2003)** | | | | |
| In a non-respondent household | **1,538** | **1,315** | No | **= 0** |
| In a respondent household | **4,931** | **6,420** | No | $W_{adjust} = \frac{W_{initial}}{\overline{R}_{RHG,\ adults}}$ |
| Cross-sectionally out-of-scope (deceased, institutionalized or outside the provinces) | **58** | **28** | No | $W_{adjust} = \frac{W_{initial}}{R_{RHG}}$ |
| **Adults aged 16 and over (on Dec. 31, 2003** | | | | |
| Cross-sectionally out-of-scope (deceased, institutionalized or outside the provinces) | **2,116** | **993** | No | $W_{adjust} = \frac{W_{initial}}{R_{RHG}}$ |
| Respondents | **25,884** | **26,957** | Yes | $W_{adjust} = \frac{W_{initial}}{R_{RHG}}$ |
| Non-respondents in a non-respondent household | **8,757** | **5,781** | Yes | **= 0** |
| Non-respondents in a respondent household (will be imputed) | **399** | **738** | No | $W_{adjust} = \frac{W_{initial}}{R_{RHG}}$ |
| **Total** | **43,683** | **4,232** | **…** | **…** |

… not applicable

$R_{RHG}$ **:** Response rate in the response homogeneity group (RHG)

$\overline{R}_{RHG,\ adults}$ **:** Average response rate of adult members of the household in which the child lives

The first method uses information collected by the LFS, that is, prior to SLID's initial interviews. The main advantage of this method is that the information is available for all longitudinal individuals, whether or not they are respondents in the current wave. The main disadvantage is that some of the variables collected are subject to change (income, level of education, marital status, employment, presence of children, etc.), and the information may be outdated for the reference year of the current wave. Thus, making an adjustment using information from wave 1 is more like adjusting the sample of respondents so that it represents the initial sample, than making a typical non-response adjustment, since the variables used may not be the ones that best explain why a household or an individual responded or did not respond. This is the method selected and used for SLID. SLID considers only the known information gathered before data collection, that is, information collected at the time of the final LFS interview.[3]

The second method involves using the most up-to-date information available for individuals. It too entails some risks. First, the information used would not necessarily be from the same period. The information about respondent longitudinal households would come from the current wave, while the information about non-respondent households would be from the previous wave. Moreover, using the latest information for non-response modelling would mean that the weight associated with the immediately preceding wave would have to be used as the base weight instead of SLID's initial weight. But that raises the problem of converted non-respondents (persons who were non-respondents in the preceding wave, who became respondents in the

3. For Panel 1, the information was collected in the preliminary interview, which was conducted at the same time as the final LFS interview (Lévesque and Franklin (2000)).

current wave). Since converted non-respondents have a longitudinal weight of zero in the wave preceding the current one, some way would have to be found to assign them a base weight.

A study was conducted to assess the possibility of changing SLID's non-response adjustment methodology (Naud, 2004). The main objective was to test the use of information from the most recent waves in non-response modelling. Two new adjustment methods were developed and evaluated. The study concluded that the use of information from more recent waves in non-response modelling had some advantages, but it was impossible to prove that such a change would clearly result in higher-quality estimates. In view of the mixed results and the complexity of the proposed methods, it was decided not to change the current methodology, and to continue studying the various ways of improving non-response modelling.

As indicated above, LFS information is used in modelling SLID non-response. Since there are a number of LFS variables, the first step is to determine which independent variables will be used in non-response modelling. This first step is not carried out until the first wave of a panel has been processed, because the variables used in the non-response model must be determined at the beginning of each panel. The variables selected will also be used in subsequent waves. To determine what explanatory variables should be used in the non-response modelling, classification variables with $m>2$ possible values are recoded as $m$ dichotomous variables. From all these variables, the ones most likely to affect response are selected subjectively.

Logistic regression is performed separately for each variable. On the basis of the results of these univariate regressions, the significant variables are selected. A total of 46 dichotomous variables were retained for Panels 3 and 4. A complete list of the variables selected for each panel is provided in Appendix A. As mentioned earlier, the same variables are used for all waves of a panel; the only changes normally made are alterations in the possible values of some variables (e.g., age group and income bracket).

The second step is to generate RHGs on the basis of the variables selected in the first step. SLID uses segmentation modelling to produce the RHGs.[4] Prior to the 2002 reference year, SLID used the KnowledgeSeeker application to generate the RHGs. Since then, the SLID team has developed a SAS program that produces the RHGs just as KnowledgeSeeker did, but in a more automated and user-friendly manner since no manual intervention is required.

Segmentation modelling employs an iterative process to partition the data file and form the decision tree. The root node of the tree is simply the entire data set. The first iteration defines the tree's first split. The latter is created by determining which of the variables selected for modelling has the most significant effect on the *response* variable. For each variable, a statistical measure reflecting the extent to which the variable discriminates non-response is computed, and the one that has the most significant effect is selected. This is how the first split is formed, with a node at each end. For each node formed, another iteration of the procedure is carried out independently. This in turn produces more splits and more nodes. When the process reaches a node where no variable has a significant effect on response, there are no further splits, and the formation of the branch is complete. The iterative process ends when all branches of the tree are completed.

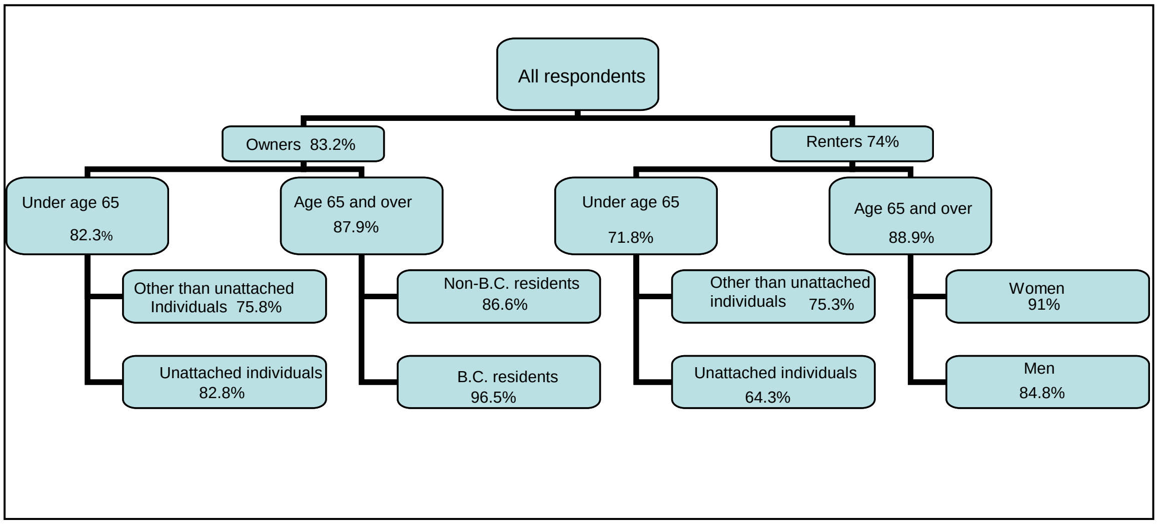
A number of methods can be used to find the variable that has the greatest influence on the tendency for response. SLID uses chi-square automatic interaction detection (CHAID), an algorithm that selects the variable with the highest Pearson chi-square statistic.

One of the reasons segmentation modelling is preferable to logistic regression is that it is more flexible. Logistic regression forces the model to be symmetric: once the significant variables are determined, the RHGs are formed by taking all possible combinations of the intersection of these variables. This process can create small cells or cells that are not significant, which can affect the reliability of the estimates. Segmentation modelling allows the model to be asymmetric: at each split in the decision tree, it determines which variable is most significant. This ensures that each RHG is significant, and it usually results in an asymmetric model. Figure 4.1 shows the key variables in discriminating between response and non-response in Panel 4 for the

4. Segmentation modelling is preferred to logistic regression because it reduces non-response bias more effectively (Dufour *et al*., 1998). However, logistic regression now appears to be favoured in some surveys, and further studies may be conducted in the near future to determine whether the segmentation modelling method should be changed.

2003 reference year. The most significant variable in this case is a variable that indicates whether the person is a homeowner.

**Figure 4.1 First few branches of the decision tree for Panel 4, 2003 reference year**




Another drawback of segmentation modelling is directly related to the decision tree itself. Once formed, a decision tree can quickly become cumbersome; beyond a certain number of splits, it becomes difficult to interpret. Decision trees are also rather unstable: the addition, removal or alteration of one variable or one individual can have a major impact on the tree. This makes it very difficult to compare several decision trees.

In each wave, SLID generates a non-response adjustment model separately for each panel. Each model is initially defined at the national level. Province-level splits are kept in the decision tree if they are significant. The advantage of a national non-response adjustment is that more variables can be used to generate the RHGs.[5]

In accordance with the sampling design, the normed initial weights are used to compute a weighted chi-square statistic. It should be noted that at this time, SLID does not distinguish between the various types of non-response (attrition, not traced, refusal, etc.)[6]. Also important is the fact that some constraints are applied to the model during segmentation modelling: each RHG must contain at least 30 people (weighted and unweighted) and have a weighted response rate of at least 50%. Kalton and Kasprzyk (1986) recommend these constraints because they prevent excessive variability in the weights and reduce loss of precision in the estimates.

Once the decision tree has been created, the RHGs are formed using the ends of the branches. The individuals used to model non-response (i.e., respondent adults and non-respondent adults in non-respondent households) are placed in the various RHGs. Note that even though non-respondent adults in respondent households are deemed respondents, they are not used in non-response modelling because they did not actually respond to the survey.

5. Studies have shown that such an approach results in differences between provincial populations before and after non-response adjustment. In particular, the populations of small provinces increase while the populations of larger provinces decline. Further studies are planned to explore this problem in greater depth and determine whether there should be a non-response adjustment at the provincial or even regional level.

6. Future studies are planned to explore the possibility of modelling non-response, while accounting for the different types of non-response.

The non-response adjustment is carried out within each RHG: the weights of the respondents in each RHG are divided by the RHG's weighted response rate (which is the same as multiplying the weights of the respondents by an adjustment factor equal to the inverse weighted response rate). Note that out-of-scope persons excluded from modelling are included in the response rate calculations for each RHG. For wave 5 of Panel 3, there were 124 RHGs, with an average weighted response rate of 71.8%. For wave 2 of Panel 4, there were 209 RHGs, with an average weighted response rate of 80.8%.

As noted above, children are not included in non-response modelling. However, their weights are adjusted for non-response to avoid creating an imbalance between households with children and households without children. The weights of children in respondent households are divided by the average weighted response rates in the RHGs of longitudinal adults living in the same household. The non-response adjustment factor for children is equal to the inverse of the average adjustment for longitudinal adults living in the same household.

Finally, the non-response weights are set to 0. Thus, during each wave, the longitudinal weights adjusted for non-response are obtained as follows

$$w_{adjust} = \begin{cases} 0, & \text{for persons in non-respondent households} \\ \dfrac{w_{initial}}{R_{GHR}}, & \text{for adults in respondent household and out-of-scope individuals} \\ \dfrac{w_{initial}}{\overline{R}_{GHR\,adults}}, & \text{for children in respondent households} \end{cases}$$

where

$w_{adjust}$ = non-response-adjusted longitudinal weight

$w_{initial}$ = initial longitudinal weight (= $w_{initial,p_3}$ for Panel 3 and $w_{initial,p_4}$ for Panel 4)

$R_{RHG}$ = weighted response rate in the RHG to which the adult belongs

$\overline{R}_{RHG,adults}$ = average weighted response rate in the RHG of adult members of the household in which the child is living

## 4.4 Adjustment for influential values

Once the weights have been adjusted to compensate for non-response, the next step is to identify influential values that have a major impact on cross-sectional income estimates and variance estimates. A value is considered influential if it has an excessive effect on the cross-sectional estimate of total provincial income. An adjustment factor between 0 and 1 is computed and applied to the weights of individuals identified as influential (and to the weights of the other members of the household) to reduce the influence their weights have on income estimates. In wave 5 of Panel 3, 10 individuals were identified as influential, and in wave 2 of Panel 4, 22 individuals were identified as influential.

Only influential values affecting the cross-sectional estimates are identified in this way. The same procedure could be followed for influential values affecting longitudinal weights. This might result in adjustments that were different for longitudinal and cross-sectional weighting. However, tests for the 1995 reference year have shown that at that time, there would have been no difference in the individuals identified as influential and very little difference in the adjustment factors. Consequently, the decision was made to use the cross-sectional adjustments for longitudinal weighting as well. For further details on the identification of influential values and the calculation of the adjustment factor, see subsection 5.6.

Some changes were made in the way the adjustment factor was calculated for the longitudinal weights. The purpose of these changes was to ensure that the weight of an influential observation in wave *t* is also adjusted for subsequent waves. This is necessary because the weight used in longitudinal analyses is often the weight in the latest available wave, and because combining that weight with income data from previous waves can have an excessive effect on the estimates. If the adjustment is applied only during processing of the wave in which the influential observation was made, the weight of an observation may not be adjusted for a specific year yet may result in an influential estimate when it is combined with income from a previous year. Consequently, the adjustment made for an influential observation is equal to the smallest of the adjustments computed for the current wave and all previous waves.

The longitudinal weight adjusted for influential values is defined as follows:

$$w_{\text{infl}} = w_{\text{adjust}} * \boldsymbol{b}_{\text{infl, min}}$$

where

- $w_{\text{infl}}$ = longitudinal weight after adjustment for influential values
- $w_{\text{adjust}}$ = non-response-adjusted longitudinal weight
- $\boldsymbol{b}_{\text{infl, min}}$ = minimum influential-value adjustment factor in the current wave and previous waves $(0 < b_{\text{infl,min}} \le 1)$

## 4.5 Longitudinal calibration

In theory, the weights obtained after adjustment for influential values could be used to produce the survey estimates. However, estimation theory tells us that if there are auxiliary data concerning the target population and those data are correlated with the characteristics being studied, they can be used to produce more accurate estimates.

There is another reason to use auxiliary data in the estimation process. It is often important that survey estimates be equal to certain known population totals. Many surveys adjust their estimates by age group and by sex to make them consistent with projected estimates based on the most recent census data. Since those population projections are revised after the next census, SLID carries out a historical revision of survey weights approximately every five years. Note that the auxiliary information used may also be derived from administrative data.

This step in the weighting process has undergone many changes since SLID was introduced. For example, in SLID's early years, a simple post-stratification process was used to adjust the weights to known population totals. The only strata used were age-sex groups at the provincial level (Lévesque and Franklin, 2000). Later, following a huge project to harmonize calibration for income and expenditure surveys and the development of new control totals, SLID underwent two major historical revisions in 2002 and 2005. New calibration strategies were developed primarily for cross-sectional weighting; they were also adapted for longitudinal weighting. For more details concerning the reasons for choosing the current cross-sectional calibration strategy, see the section on cross-sectional calibration (5.7).

The control totals used in calibration are chosen on the basis of their availability and quality. Administrative data are often subject to error, and it is important to ensure that the adjustment is based on data of reasonable quality. Hence, even if the control totals used also contain errors, the errors must be smaller than the error that would result from not adjusting the survey estimates. Another constraint on the choice of control totals is that the corresponding variables in the survey must be available for all respondents.

Table 4.3 shows the history of the control totals used in SLID weighting from the introduction of the survey to the most recent historical revision in 2005.

**Table 4.3 History of control totals used in calibration**

| Control totals | Before revision[1] (post-stratification) | 2002 historical revision (2000 RY) | 2005 historical revision (2003 RY) |
|---|---|---|---|
| **Age-sex groups** | 0 to 6 45 to 54<br>7 to 15 55 to 59<br>16 to 18 60 to 64<br>19 to 24 65 to 69<br>25 to 34 70 and<br>35 to 44 over | 0 to 6 45 to 54<br>7 to 15 55 to 59<br>16 to 17 60 to 64<br>18 to 24 65 to 69<br>25 to 34 70 and<br>35 to 44 over | 0 to 6 (both sexes) 45 to 54<br>7 to 17 55 to 64<br>18 to 24 65 and over<br>25 to 34<br>35 to 44 |
| **Number of households by size[2]** | … | size = 1, size = 2 | size = 1, size = 2 |
| **Number of economic entities by size[3]** | … | size = 1, size = 2 | size = 1, size = 2 |
| **Number of paid employees in each wages and salaries class** | … | … | 7 classes defined by 10th, 25th, 50th, 65th, 75th and 95th, 98th or 99th percentile |
| **Period affected by the revision** | **…** | **1980 to 2000** | **1990 to 2003** |

… not applicable
1. For Panel 1, other age groups were used in post-stratification (Lévesque and Franklin, 2000).
2. A household consists of a person or group of persons who occupy the same dwelling.
3. An economic entity (also known as an economic family) consists of 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. In this report, an economic entity of size 1 (also referred to as an unattached individual) is a person living alone or with one or more other persons who are not related by blood, marriage, common-law or adoption.

### 4.5.1 Calibration theory

SLID uses a generalized regression (GREG) estimator in the calibration process (Särndal *et al.*, 1992). The GREG estimator adjusts survey estimates to known population totals (control totals) while minimizing the distance between weights before and after calibration, and taking into account certain constraints on the final weights. It is a model-assisted estimation method that is capable of calibrating with a number of variables simultaneously. The restriction on SLID's final weights is that they must be greater than or equal to 1 at the end of the calibration process.

Let *Y* be the variable of interest and $\underline{y}$ be the *1* x *n* vector of values in the sample. The notation used is similar to the notation used by Bankier (2002). Let $\underline{\hat{x}}$ be the *p* x *n* matrix of values in the sample for the *p* auxiliary variables (or control totals). The estimator takes the form

$$\hat{Y} = \underline{y}\ diag(\underline{w})\ \underline{g}$$

where $\underline{w}$ is an *n* x *1* vector of pre-calibration weights, $diag(\underline{w})$ is an *n* x *n* diagonal matrix with the $\underline{w}$ pre-calibration weights on the diagonal, and $\underline{g}$ is the *n* x *1* vector of g-weights. The g-weights are calculated so that that the loss function *L* is minimized under the constraint

$$\underline{x}\, diag(\underline{w})\, \underline{g} = \underline{X}$$

where $\underline{X}$ is the *p* x *1* vector of the *p* control totals. The loss function used in SLID is

$$L = (\underline{g} - \underline{1}_n)^{/}\ \underline{I}_n\, (\underline{g} - \underline{1}_n)$$

where $\underline{I}_n$ is the identity matrix. The resulting g-weights are given by

$$\underline{g} = \underline{1}_n + \underline{\hat{x}}^{/}\, (\underline{\hat{x}}\underline{\hat{x}}^{/})^{-1}\, (\underline{X} - \underline{\hat{x}}^{/}\underline{1}_n)\,.$$

### 4.5.2 Applying Calibration to the Longitudinal Weights

In longitudinal weighting, calibration is performed independently for each panel, since the survey estimates have to be adjusted so that they represent the target population at the time the panel was selected (December 31 of 1998 and 2001 for Panels 3 and 4 respectively).

Individuals are assigned to the various calibration classes using the characteristics of the longitudinal individuals at the time the panel was selected (wave 0). Thus, the available information about all respondents comes from the Labour Force Survey (LFS). For example, the province used in longitudinal calibration is the province in which the individual lived at the time the sample for a given panel was selected, and the age group to which an individual belongs is determined by his/her age at the time the sample for the panel concerned was selected.

As noted above, the current calibration strategy was originally developed for the cross-sectional part of the survey. Table 4.3 shows the control totals selected when the most recent calibration strategy was introduced. It contains age-sex groups, number of households and economic entities by size, and number of paid employees by wages and salaries class.

The control totals selected for the cross-sectional part could not be used directly for the longitudinal part. Longitudinal individuals cannot be assigned to wages and salaries classes in the initial wave because that information is not available in the LFS. Table 4.4 shows the totals and variables that are available in the various waves for longitudinal calibration.

**Table.4.4 Control totals and survey variables available for longitudinal calibration**

| Availability for calibration | | Wave 0 | Wave 1 | Wave X |
|---|---|---|---|---|
| Survey variables | Demographic | ✔ | ✔ | ✔ |
| | Wages and salaries | … | ✔ | ✔ |
| Control totals | Demographic | ✔ | ✔ | ✔ |
| | Wages and salaries | ✔ | ✔ | ✔ |

**Target Population**

Several options for circumventing the problem caused by the unavailability of the "wages and salaries" survey variable in wave 0 were suggested and evaluated. The solution selected was to perform a two-step calibration using available information from wave 1. Step 1 involves adjusting the survey estimates corresponding to wave 1 to the control totals for the same period. All control totals are used in step 1. Step 2 involves adjusting the survey estimates so that they represent the target population (the wave 0 population). Only the demographic control totals are used, since they are available in the LFS. Because the demographic totals change little between consecutive years, completing the second calibration without the wages and salaries classes has little impact on the resulting wages and salaries distribution. The second calibration adjusts the weights so that the data represent the initial population while maintaining an acceptable wages and salaries distribution.

**Table 4.5 Example of control totals used to calibrate panel 4**

| | | Control Totals | |
|---|---|---|---|
| **First calibration**<br>Wave 1<br>Reference period:<br>December 31, 2002 | Age-sex group | 1 to 7 (both sexes)<br>8 to 18<br>19 to 25<br>26 to 35 | 36 to 45<br>46 to 55<br>56 to 65<br>66 and over |
| | Number of households by size | Size=1 and size=2 | |
| | Number of economic entities by size | Size=1 and size=2 | |
| | Wages and salaries classes | 7 classes defined by 10th, 25th, 50th, 65th, 75th and 95th, 98th or 99th percentile | |
| **Second calibration**<br>Wave 0<br>Reference period:<br>December 31, 2001 | Age-sex group | 0 to 6 (both sexes)<br>7 to 17<br>18 to 24<br>25 to 34 | 35 to 44<br>45 to 54<br>55 to 64<br>65 and over |
| | Number of households by size | Size=1 and size=2 | |
| | Number of economic entities by size | Size=1 and size=2 | |

On the basis of the control totals shown in Table 4.5 and the theory outlined in section 4.5.1, the vector of the longitudinal weights produced by the first calibration is defined as

$$\underline{\mathrm{w}}_{\mathrm{cal},1} = \mathrm{diag}\left(\underline{\mathrm{w}}_{infl}\right)\underline{\mathrm{g}}_1$$

where $\mathrm{diag}\left(\underline{\mathrm{w}}_{infl}\right)$ is an $n$ x $n$ diagonal matrix with the $\underline{\mathrm{w}}_{infl}$ pre-calibration weights (i.e., the weights after adjustment for influential values) on the diagonal, and $\underline{\mathrm{g}}_1$ is the vector of the $n$ x $1$ g-weights produced by the first calibration.

The longitudinal weights vector produced by the second calibration is given by

$$\underline{\mathrm{w}}_{\mathrm{cal},2} = \mathrm{diag}\left(\underline{\mathrm{w}}_{\mathrm{cal},1}\right)\underline{\mathrm{g}}_2$$

where $\mathrm{diag}\left(\underline{\mathrm{w}}_{\mathrm{cal},1}\right)$ is a diagonal $n$ x $n$ matrix with the weights resulting from the first calibration, $\underline{\mathrm{w}}_{\mathrm{cal},1}$, on the diagonal, and $\underline{g}_2$ is the $n$ x $1$ vector of the g-weights produced by the second calibration.

The average adjustment in the longitudinal weights adjusted for non-response and influential values following the calibration process is 1.43 for wave 5 of Panel 3 and 1.4 for wave 2 of Panel 4.

The set of weights resulting from the two-step calibration process is the set of final longitudinal weights for SLID.[7] Table 4.6 shows the median final longitudinal weight for each province for the 2003 reference year. The province shown in the table is the individual's province of residence on December 31 of the reference year (2003). Individuals who were deceased, institutionalized or outside Canada's ten provinces on December 31, 2003, are included in a separate category.

7. At one time, SLID also produced a set of weights with random noise added for external use (Lévesque and Franklin, 2000). Those weights are no longer produced.

**Table 4.6 Median final longitudinal weights for wave 5 of panel 3 and wave 2 of panel 4, 2003 reference year**

| Province | Final longitudinal | |
|---|---|---|
| | Panel 3 | Panel 4 |
| | weight | |
| Newfoundland and Labrador | 312 | 331 |
| Prince Edward Island | 131 | 134 |
| Nova Scotia | 364 | 354 |
| New Brunswick | 331 | 351 |
| Quebec | 794 | 727 |
| Ontario | 833 | 781 |
| Manitoba | 425 | 448 |
| Saskatchewan | 358 | 354 |
| Alberta | 1,004 | 1,065 |
| British Columbia | 1,283 | 1,129 |
| Deceased, institutionalized or outside the provinces | 601 | 680 |

## 5 Cross-sectional weighting

This section provides a detailed description of the steps required to create a set of weights representative of Canada's ten provinces on December 31 of a given reference year. All cross-sectionally in-scope longitudinal individuals and their cohabitants are included in the production of this set of weights. The methodology described here is the one used for the 2003 reference year.

SLID produces two types of cross-sectional weights: integrated weights and individual weights. The difference between them is that the individual weight can be different for each member of a household, while the integrated weight is the same for all members of a household. Integrated weights are suitable for studies at the household or family level, but they can also be used for person-level studies. Individual weights, on the other hand, can only be used at the person level.[8]

The first set of weights produced – the *integrated weights* – is the set most commonly used to compute the cross-sectional estimates included in SLID publications. The second set of weights produced is the *labour weights* which are individual weights. Integrated weights and labour weights differ somewhat in the definition of respondent they use. As the name suggests, the labour weight uses only the labour interview responses, and not the responses to both the labour and income interviews, to identify a respondent. For example, a household that is non-respondent on the labour interview (i.e., a household whose members are all non-respondents on the labour interview) but respondent on the income interview (because permission to access tax data was obtained previously) will be assigned a labour weight of zero in the weighting process. Using labour weights rather than integrated weights provides a better estimate of labour characteristics.

The cross-sectional weighting process involves the following seven steps: determination of individuals eligible for cross-sectional weighting (subsection 5.1), non-response adjustment (5.2), interprovincial migration adjustment (5.3), application of panel allocation factors (5.4), weight share (5.5), adjustment for influential values (5.6) and calibration (5.7).

### 5.1 Determination of individuals eligible for cross-sectional weighting

The cross-sectional sample for a given year is representative of the population of Canada's ten provinces on December 31 of the reference year (except for households composed entirely of immigrants). It consists of longitudinal individuals who are still members of the population of Canada's ten provinces as of December 31 of the reference year, plus their cohabitants. These individuals are considered to be *cross-sectionally eligible*.

Before we can determine which individuals have a non-zero cross-sectional weight, we must first identify those who are cross-sectionally eligible (in-scope). All longitudinal individuals who are no longer members of the target population on December 31 of the reference year (e.g., persons who have moved to a place outside the ten Canadian provinces, deceased persons, and persons living in an institution) are assigned a cross-sectional weight of zero, even if they are respondent (or deemed respondent). Of the longitudinal individuals who belong to the target population on December 31 of the reference year, only those who are members of respondent households have non-zero cross-sectional weights. For the purposes of producing the integrated weights, the concept of a respondent household is as follows. A household is defined as respondent if at least one member responded in at least one of the two interviews (the labour interview or the income interview). It was also decided that a household's composition cannot be missing for more than a year. For example, if a household is non-respondent on the labour interview for two consecutive years, it will be considered non-respondent. This condition was added during the 2003 historical revision to avoid having large numbers of respondent households for which there was no recent information about household composition (in particular, it eliminates cases for which there is permission to match income tax files but no other information).

8. Previously, SLID produced some sets of weights for internal use and others for external use. However, the external weights are no longer generated, and only the internal weights are used (Lévesque and Franklin, 2000).

Information about non-respondent persons in respondent households will be imputed. For the purposes of labour weights, a household is respondent if at least one member responded in the labour interview.

## 5.2 Non-response adjustment

As noted above, the cross-sectional sample for a given year is representative of the population of Canada's ten provinces on December 31 of the reference year. Like any other survey, however, SLID is subject to non-response (in this case, we are referring to total non-response), which affects how representative the sample can be. To mitigate the effects of non-response bias, we adjust the weights of respondents to compensate for non-respondents.

From a cross-sectional standpoint, we have too little information about non-respondent cohabitants to adjust for non-response at that level. For many cohabitants, the only available variables are province, age and sex, which in any case will be used in the calibration step. To circumvent this problem, two methods of cross-sectional non-response adjustment have been proposed (Latouche and Michaud, 1997). One solution is to first make a non-response adjustment for longitudinal individuals only, and then to use the weight share method to assign a weight to cohabitants. The weights of cohabitants are thus indirectly adjusted for non-response. The other method is to adjust for non-response at the household level using the initial dwelling strata and rotation groups of the LFS sample design. However, household-level adjustment tends to underestimate the response rates slightly.

SLID adopted the first method described above. Consequently, the cross-sectional non-response adjustment step is not related to cross-sectional weighting but to longitudinal weighting. Longitudinal non-response modelling and adjustment were described in subsection 4.3 above. As indicated there, for longitudinal weighting, each panel is treated separately. The cross-sectional weight adjusted for longitudinal non-response is therefore equal to the weight $w_{adjust}$ derived in subsection 4.3. We denote the cross-sectional weight adjusted for longitudinal non-response as $w_{2003,adjust}$ . This weight is used as the initial weight for the production of all derived cross-sectional weights (integrated weights and labour weights).

As noted earlier, the cross-sectional non-respondents group is of little consequence here since the non-response adjustment is made during longitudinal weighting. Only longitudinal households that are deemed to be respondent will have a non-zero weight at this stage, and cohabitants of those households will be considered in the subsequent cross-sectional weighting process.

## 5.3 Adjustment for interprovincial migration

The adjustment for interprovincial migration is another step that is a direct result of the longitudinal nature of SLID. Over time, some people in the sample move from one province to another. Since the probability of selection can vary widely with geographical location, there is a substantial difference at the national level between the largest weights and the smallest ones. As a result, if a person with a fairly low selection probability (and therefore a fairly large weight) moves to a location where a given individual's selection probability is much higher (and that individual's weight is much smaller), problems can arise. In particular, certain cases of interprovincial migration can distort the results of analytical studies.

A simple example will help explain the problem. Suppose a heart surgeon moves from Toronto to Charlottetown. This would cause several difficulties. First, the surgeon's weight of about 1,100 would be much larger than any of the weights in Prince Edward Island, which average close to 65. Even though he would represent only 0.7% of the province's population, his weight combined with his high income would artificially inflate the total income of individuals in Prince Edward Island. This is unacceptable from an analytic perspective.

A number of options are available to alleviate the problem of outlier weights due to interprovincial migration. Such weights could be replaced with the average weight, the median weight, a quantile of the weight distribution in the province of destination, or some other value based on that distribution. The alternatives are arbitrary, and any of them could be used. What is needed is a compromise between assigning the migrant a weight similar to the average weight in the province of destination and excessively altering the migrant's original weight (based on selection probability). We must first identify those persons who have migrated to other provinces and whose weights are causing an imbalance in the estimates for their provinces of destination, and then adjust the weights.

The method adopted by SLID uses the 95$^{th}$ percentile of the distribution of weights. This is a quick, simple method that requires little or no manual intervention. First it is necessary to identify every person who has moved to another province and whose weight is larger than the largest weight of the province's current population. The weights of such persons are then set to the 95$^{th}$ percentile of the province's weight distribution. Next, a random number between 0.00 and 10.00 is subtracted from that percentile to ensure that the weights remain nearly unique (two people can have the same individual weight, but such cases are very rare).

In the most recent historical revision, changes were made in this part of the weighting process. Previously, adjustments were made after the weight share step, independently for individual and integrated weights (see Lévesque and Franklin, 2000). The adjustment method was different because in the case of the integrated weights, it was important to keep the weights in a household equal.

The new method is applied after the non-response adjustment. This ensures that only one type of adjustment is made at the individual level for all sets of cross-sectional weights. Because this step is carried out at a different point in the process, interprovincial migration adjustments are applied only to longitudinal individuals on an individual basis. Thus, only longitudinal individuals who move to another province and whose non-response-adjusted weight exceeds the maximum weight for the province of destination will have their weights adjusted for interprovincial migration. Since the weights of these individuals are subsequently shared with their cohabitants, the adjustment will indirectly affect the other members of the household.

Interprovincial migration adjustments are performed and rederived each processing year. The cross-sectional weight adjusted for interprovincial migration for the 2003 reference year is defined as

$$w_{2003,\mathrm{mig}} = w_{2003,\mathrm{adjust}} \cdot a_{\mathrm{mig}}$$

where

$w_{2003,\mathrm{mig}}$ = cross-sectional weight adjusted for interprovincial migration

$w_{2003,\mathrm{adjust}}$ = cross-sectional weight adjusted for non-response

$a_{\mathrm{mig}}$ = interprovincial migration adjustment factor

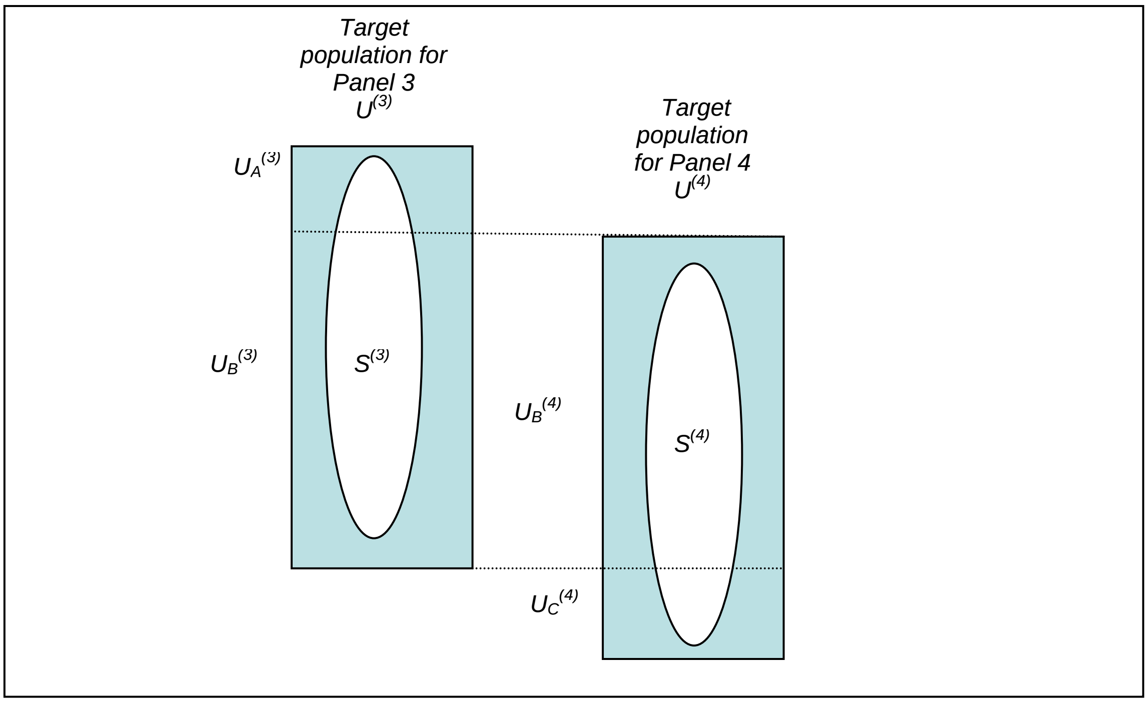
For the 2003 reference year, an interprovincial migration adjustment was needed for 94 longitudinal persons in Panel 3 and 34 in Panel 4. More interprovincial migration adjustments were made for Panel 3 than for Panel 4 because the 2003 reference year corresponds to wave 5 of Panel 3 and only wave 2 of Panel 4.

## 5.4 Application of panel allocation factors

The 1996 reference year marked the beginning of overlapping panels in SLID. In theory, the two panels could be used independently of one another to produce cross-sectional estimates, but in practice, it is better to combine the two cross-sectional samples to produce the cross-sectional estimates. This doubles the sample size, which in turn reduces the variability of the estimates and permits the use of more control totals in the calibration step.

Since the two longitudinal samples are representative of the population of Canada's ten provinces at different reference periods (December 31, 1998, for Panel 3, and December 31, 2001, for Panel 4), a way has to be found to combine the two samples so that they both represent the target population for a given reference year. The problem that arises is the time lapse between them. Figure 5.1 shows how the population of the ten provinces changed between the selection of Panel 3 and 4 of SLID.

**Figure 5.1 Change in the population of Canada's ten provinces between the selection of panels 3 and 4**




Sample $S^{(3)}$ was selected from the population of Canada's ten provinces $U^{(3)}$ in January 1999. (This was the longitudinal sample for Panel 3.) Sample $S^{(4)}$ was selected from the population of Canada's ten provinces $U^{(4)}$ in January 2002 to form the longitudinal sample of Panel 4. Most individuals present in the January 1999 population are also present in the January 2002 population ($U_B^{(3)} = U_B^{(4)}$). The differences between populations $U^{(3)}$ and $U^{(4)}$ are due to departures from the target population between January 1999 and January 2002 (deaths, persons placed in institutions, persons who moved to places outside the provinces, etc.) and to additions to the population after January 1999 (immigrants, births). Part of the Panel 3 sample is composed of persons who left the population between January 1999 and January 2002 ($U_A^{(3)}$); since those individuals were no longer members of the Canadian population when the sample for Panel 4 was selected, they had no chance of being selected. Conversely, part of the Panel 4 sample is made up of persons who were in the population of Canada's ten provinces when that sample was selected but were not in the population when Panel 3 was selected ($U_C^{(4)}$); those persons had no chance of being selected for Panel 3.

To develop a method of combining the information from the cross-sectional samples of the two panels, we need only consider the two panels at the time Panel 4 was introduced. The population of interest is $U^{(4)}$, the target population at the time Panel 4 was introduced. The longitudinal samples of interest are the original samples selected, with the Panel 3 sample having been reduced by population departures and non-respondents. SLID uses Merkouris's method (1999) to match overlapping samples (in this case, the samples for Panels 3 and 4). If Y is a population parameter, we can use the following combined estimator (which assigns a relative importance to each panel based on a certain factor $p$ ):

$$\hat{Y} = p\,\hat{Y}_3 + (1 - p)\,\hat{Y}_4$$

where $\hat{Y}$ is the cross-sectional estimate for a variable of interest, $\hat{Y}_i$ is the estimate produced for that variable for Panel *i*, and $p$ is the panel allocation factor. This estimator can be regarded as a weighted sum of the estimators of the two panels, where the weights are the panel allocation factors. The optimum value of the panel allocation factor was set so as to minimize the variance of $\hat{Y}$ . The panel allocation factor is defined as follows:

$$p = \left(\frac{n_1}{d_1}\right)\left(\frac{n_1}{d_1} + \frac{n_2}{d_2}\right)^{-1} = \frac{n_1}{n_1 + n_2\left(\frac{d_1}{d_2}\right)} \tag{5.1}$$

where $d_i$ denotes the design effect for Panel *i* and $n_i$ represents the number of Panel *i* longitudinal persons aged 16 and over who were respondent at time *t* (where *t* is the reference year concerned, which in this case is 2003). The design effect is defined as the ratio of the variance of the subweighted estimate (i.e. without benchmarking to population control totals) to the variance associated with a simple random sample of the same size. For SLID, the number of longitudinal persons aged 16 and over who were respondent at time *t* is the same as the number of cross-sectionally in-scope longitudinal persons aged 16 and over who are in a respondent household (since all members of respondent households are deemed respondent (see subsection 4.2)). Even though cohabitants are used to produce the cross-sectional estimates, only longitudinal persons respondent at time *t* are considered in computing the panel allocation factors. Merkouris (1999) and Lavallée (1994) suggest combining the two panels before incorporating cohabitants by the weight share method. Leaving cohabitants out of the computation of factor *p* helps to reduce the importance of the older panel and thus to minimize the potential effects of attrition bias.

To determine the optimum factor *p*, we need to estimate the design effects and the longitudinal sample sizes. However, no design effect was available at the time of data processing. Since SLID is a subsample of the LFS, it was decided to use LFS data to derive the design effects needed to compute the *p* factors. Moreover, since we are interested in the ratio of two design effects, the LFS is a perfectly legitimate source for the required effects. The LFS's six rotation groups are used to determine the design effects. The effects are calculated for the *number of persons aged 15 and over* characteristic at the provincial level (which will require the calculation of ten panel allocation factors, one per province).

For Panels 1 and 2 combined, the LFS files for January 1993 and January 1996 were used to compute the design effects (Lévesque and Franklin, 2000). The two panels were from different sampling plans since the LFS underwent a postcensal redesign between the selection of Panel 1 and the selection of Panel 2 (completed at the end of 1994). For Panels 3 and 4, however, it was decided not to use the design effects in computing the allocation factors because both panels were selected under the same sampling plan and the LFS design effects were very similar. Hence, the design effect ratio was 1. The actual ratio, however tends to favour the older panel.

The required sample sizes were computed from the longitudinal samples of the two panels for the reference year concerned. As noted above, the sample sizes are the number of longitudinal persons aged 16 and over who were respondent at time *t*. They are obtained by counting all cross-sectionally in-scope persons aged 16 and over in respondent households.

Once the design effects and the sample sizes are known, it is easy to compute the panel allocation factors using equation 5.1. The factors must be calculated for each year involved since the composition of the longitudinal samples changes every year. For practical reasons, it was decided to compute the design effects at the inception of each new panel and to use those design effects in calculating the panel allocation factors for the three overlap years (Latouche *et al*., 2000). The panel allocation factors for the 2003 reference year are presented in Table 5.1 below.

**Table 5.1 Panel allocation factors for the 2003 reference year**

| Province | $d_3/d_4$ [1] | $n_3$ | $N_4$ | $P$ Panel 3 | $(1 - P)$ Panel 4 |
|---|---|---|---|---|---|
| | | | percentage | | |
| **Newfoundland and Labrador** | 1 | 1,214 | 1,139 | 0.5159 | 0.4841 |
| **Prince Edward Island** | 1 | 751 | 820 | 0.4780 | 0.5220 |
| **Nova Scotia** | 1 | 1,793 | 1,822 | 0.4960 | 0.5040 |
| **New Brunswick** | 1 | 1,620 | 1,571 | 0.5077 | 0.4923 |
| **Quebec** | 1 | 4,972 | 5,386 | 0.4800 | 0.5200 |
| **Ontario** | 1 | 7,570 | 8,288 | 0.4774 | 0.5226 |
| **Manitoba** | 1 | 1,990 | 1,961 | 0.5037 | 0.4963 |
| **Saskatchewan** | 1 | 1,924 | 1,923 | 0.5001 | 0.4999 |
| **Alberta** | 1 | 2,197 | 2,246 | 0.4945 | 0.5055 |
| **British Columbia** | 1 | 2,252 | 2,539 | 0.4700 | 0.5300 |

1. For the combination of Panels 3 and 4, a design effect ratio of 1 was used since there was little difference between the LFS design effects for the two periods.

$d_i$ = design effect for Panel *i*

$n_i$ = number of Panel *i* respondent longitudinal persons aged 16 and over for the 2003 reference year

$p$, $1 - p$ = panel allocation factors

The table shows that the panel allocation factors are close to 0.5, and that the two panels have similar importance; this contrasts with what was done in the calculation of allocation factors between Panels 1 and 2 (see Lévesque and Franklin, 2000).

The *p* factors are applied to all Panel 3 longitudinal persons who were in a respondent household at time *t* (in this case, December 31, 2003) and who were members of the target population at the time Panel 4 was selected. Thus, the *p* factor is not applied to Panel 3 persons who were institutionalized or lived outside the ten provinces at the time Panel 4 was selected, since they had no chance of being selected for the Panel 4 sample. Prior to the most recent historical revision, the *p* factor was applied to all members of the older panel (Lévesque and Franklin, 2000). This change was implemented retroactively for all SLID years.

The *1-p* factors should be applied to all Panel 4 longitudinal persons who were in a respondent household at time *t*, except persons who were outside the target population when Panel 3 was introduced (January 1999) and in the target population when Panel 4 was introduced (January 2002). For practical reasons, only newborns and immigrants[9] who were not present when Panel 3 was introduced, but were present when Panel 4 was introduced, are exempted from application of the *1-p* factors. Those excluded will retain their weights as adjusted for interprovincial migration. Accordingly, the cross-sectional weight adjusted for the combination of the two panels is as follows:

9. Questions on immigration are only asked of people aged 16 and up as of the January interview. Thus, the immigration status and the year of immigration are imputed based on the information of the closest parent, in order to identify any children born before the older panel was selected, and who immigrated to one of the ten Canadian provinces during the period included between the two panels. Once the information for a child has been obtained (i.e. once he or she is 16 years old), it replaces the previously imputed values.

$$w_{2003,p} = \begin{cases} p\, w_{2003,\mathrm{mig}}, & \text{for Panel 3 persons not exempted from application of the panel allocation factors,} \\ (1 - p)\, w_{2003,1} & \text{for Panel 4 persons not exempted from application of the panel allocation factors,} \\ w_{2003,\mathrm{mig}}, & \text{For Panel 3 and Panel 4 persons exempted from application of the panel allocation factors.} \end{cases}$$

where $w_{2003,p}$ = Cross-sectional weight adjusted for the combination of the two panels
$w_{2003,\mathrm{mig}}$ = Cross-sectional weight adjusted for interprovincial migration
$p$ = Panel allocation factor

For the integrated weight and the labour weight, the weight adjusted for the combination of the panels is computed in the same way, but the concept of a respondent household is different in the case of the labour weight (see subsection 5.1). Note that the panel allocation factors are applied to the weights after the non-response and interprovincial migration adjustment step and before the weight share step, so that they will be included in the weights of cohabitants.

## 5.5 Weight share

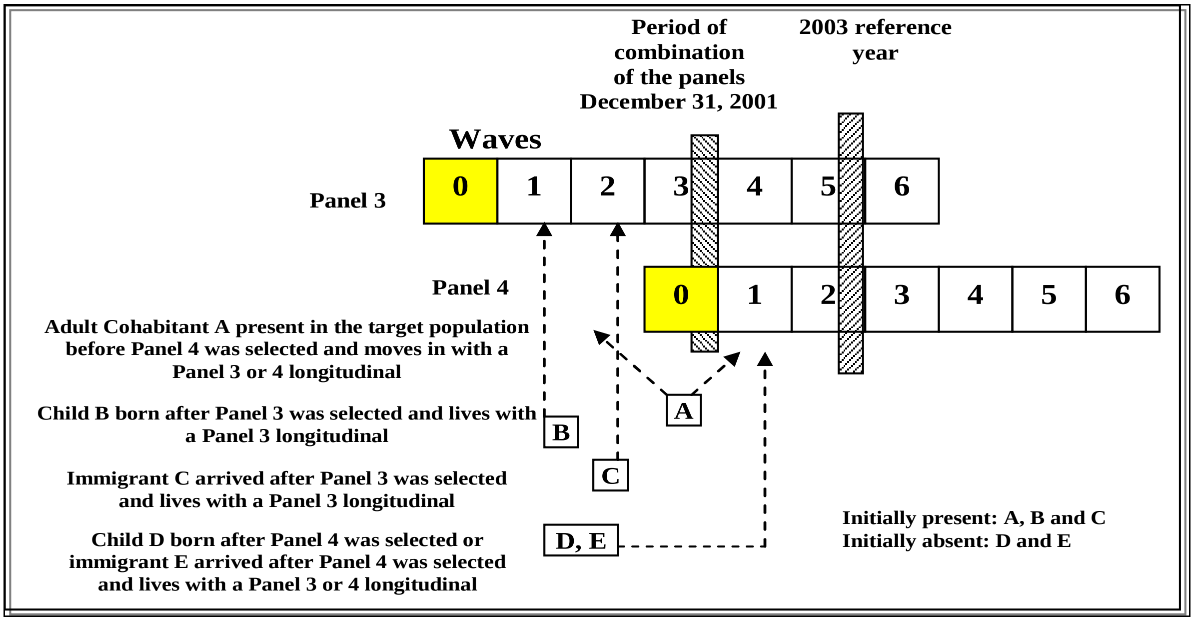
The weight share step is unique and essential to cross-sectional weighting. It is necessitated by the longitudinal nature of SLID. In a longitudinal survey, the sample evolves over time. New people move in with the selected individuals, and children are born into some of the sample households. In short, persons who were not originally selected become part of the sample. Since we do not have any probability of selection for these new persons (cohabitants), we have to find a way to assign them a weight. This is done by the weight share method. A brief explanation of this method is presented below. For further details, the reader should refer to the article by Lavallée (1995) and Lavallée (2002).

It must first be determined at what level in the sample design hierarchy the weight share is to be performed: province, stratum, primary sampling unit or final sampling unit. SLID has elected to perform the weight share at the level of the final sampling unit, the household. Once that decision is made, it is necessary to determine the status of each person in the cross-sectional sample: longitudinal individual, initially present cohabitant or initially absent cohabitant.

From the weight share step onward, no distinction is made between the two panels, and all individuals are treated as part of the same sample, though the concepts of longitudinal individual and cohabitant remain. One might think that since we choose to perform the weight share at the household level, there would be no difference between treating the two cross-sectional samples separately or together. It should be considered that in a survey with multiple overlapping panels, some households may contain persons from the oldest panel living with persons from the most recent panel. When the longitudinal sample for Panel 4 was selected, any household containing at least one longitudinal person from Panel 3 was excluded in order to avoid surveying the same household for another six years. Because of the survey's dynamic, longitudinal nature, however, over a six-year period households may break up and new households may form. As a result, there is a possibility that the survey will include households with members from the two panels. If the cross-sectional samples are considered separately, two distinct sets of weights would be produced for such households, and this would cause all sorts of problems. For this reason, it is better to consider the two panels together.

A cohabitant is initially present if he or she was a member of the target population when Panel 4 was selected in January 2002 (Adult Cohabitant A, Child B and Immigrant C in Figure 5.2). Conversely, a cohabitant is initially absent if he or she was not a member of the target population when Panel 4 was selected (Persons D and E in Figure 5.2). Since this discussion is about the overlap of two panels, the concepts of initially absent and initially present cohabitant must reflect the fact that only one panel is being considered and that the two panels must be considered together. The two panels must be considered at the time that Panel 4 was selected (January 2002). It is important to note that all cohabitants have a longitudinal weight of zero, since they are not part of the longitudinal sample. They may, however, have a non-zero cross-sectional weight, assigned by the weight share method.

**Figure 5.2 Definition of initially absent and present cohabitants for the 2003 reference year**




In the weight share step, two different weights are produced for each individual: the integrated weight (which is the same for all members of a given household) and the labour weight, which is an individual weight (and may be different for each member of a given household). The way in which the shared weights are calculated depends on whether there are cohabitants initially present in the household and/or initially absent, and on whether the weight is an integrated weight or an individual weight.

### 5.5.1 Weight share for integrated weights

In the case of the weight share for integrated weights, the weights resulting from the share must be identical for all members of the household. This is necessary because the pre-calibration weight must be identical for all members of a household so that an integrated weight can be produced (Lemaître and Dufour, 1987). However, the weights adjusted for the combination of the panels are not identical for all members of a household, since previous adjustments are made mostly at the individual level. For a given household *h*, the integrated shared cross-sectional weight is therefore defined as

$$w_{2003,\,\text{shared}} = \frac{\sum_h w_{2003,\,p}}{n_{L,\,h} + n_{IP,\,h}}$$

where $w_{2003,\,\text{shared}}$ = shared cross-sectional weight

$w_{2003,\,p}$ = cross-sectional weight adjusted for the combination of the two panels

$n_{L,\,h}$ = number of longitudinal persons in household *h*

$n_{IP,\,h}$ = number of cohabitants initially present in household *h*

### 5.5.2 Weight share for labour weights

Since the labour weight is an individual or non-integrated weight, the weight share process involves the following steps. First, the households to which cohabitants have moved since the longitudinal sample was selected must be identified. For households composed entirely of longitudinal persons, the weight share method is not needed. For all members of such households, the shared cross-sectional weight is identical to the adjusted cross-sectional weight after the panel allocation factors are applied.

For households with new cohabitants, the cross-sectional weights adjusted for the panel allocation factor $w_{2003,p}$ of all longitudinal household members must be added together. Then the longitudinal individuals and the initially present cohabitants in each household must be counted.

For households with at least one cohabitant initially present, the methodology was changed at the time of the most recent historical revision to preserve the longitudinals' non-response adjustment and the ratio between the weights of the household's longitudinals. For more details on the old methodology, see Lévesque and Franklin (2000). The shared weight is now calculated differently for longitudinal individuals and for cohabitants. For cohabitants, the shared cross-sectional weight is the sum of the cross-sectional weights $w_{2003,p}$ of all longitudinal household members divided by the total number of longitudinal individuals and initially present cohabitants in the household. For longitudinal individuals, the shared cross-sectional weight is the pre-share weight $w_{2003,p}$ multiplied by a factor equal to the number of longitudinal individuals divided by the sum of the longitudinal individuals and initially present cohabitants.

For households with all cohabitants initially absent, the shared cross-sectional weight of longitudinal individuals and cohabitants is calculated in a different way. For longitudinal members of such households, the shared cross-sectional weight is identical to the adjusted cross-sectional weight after the panel allocation factors are applied. For cohabitants, the shared cross-sectional weight is the sum of the weights $w_{2003,p}$ of all longitudinal household members divided by the total number of longitudinal individuals. Thus, the longitudinal individuals will not be assigned an average weight, and the non-response adjustment will be preserved. Only initially absent individuals will be assigned the average weight of the household's longitudinal members.

The shared individual cross-sectional weight (usual individual weight or labour weight) is defined as

$$w_{2003,\text{shared}} = \begin{cases} w_{2003,p}, & \text{for persons in households composed entirely of longitudinal persons, and for longitudinal members of households in which all cohabitants are IA} \\ w_{2003,p} \cdot \left( \dfrac{n_{L,h}}{n_{L,h} + n_{IP,h}} \right), & \text{for longitudinal persons in households composed of longitudinal persons and at least one IP cohabitant} \\ \dfrac{\sum_h w_{2003,p}}{n_{L,\,h} + n_{IP,\,h}}, & \text{for all cohabitants} \end{cases}$$

where $w_{2003,\,shared}$ = shared cross-sectional weight

$w_{2003,\,p}$ = cross-sectional weight adjusted for the combination of the two panels

*IP* = initially present

*IA* = initially absent

$n_{L,\,h}$ = number of longitudinal persons in household *h*

$n_{IP,\,h}$ = number of cohabitants initially present in household *h*

For the 2003 reference year, 75,206 individuals have non-zero integrated cross-sectional weights, and 70,278 have non-zero labour cross-sectional weights.

### 5.6 Adjustment for influential values

Like any survey that attempts to estimate income distribution, SLID is very likely to face the problem of influential values (such distributions have very long tails). Influential values have a significant impact on provincial estimates of total and average income and their variance estimates. The impact is even greater on estimates for smaller domains such as age group, sex or family size. Hence, methods must be developed to identify the influential values and then adjust them so as to reduce the effect they have on estimates.

It should be noted that the identification of influential values and calculation of adjustment factors steps are performed only on the integrated weights, even though the derived adjustment factors are equally applied to both the integrated weights and labour weights. Before the last historic revision, this step was performed using the individual weight (Lévesque and Franklin, 2000).

SCF devised a method of detecting and processing influential values (Tremblay, 1998). Since it was known that SCF would be incorporated into SLID for the 1999 collection period, the methodology takes the fact that the successor survey is longitudinal into account. Accordingly, SLID adopted this approach to deal with the problem of influential incomes. The first step is to identify the influential items, that is, observations that contribute too heavily to the weighted estimate of total personal income for the province. Such observations usually occur in upper income brackets. An individual's income (in this case, total income minus capital gains) is regarded as influential if the contribution of the individual's weighted income to the provincial weighted estimate exceeds a certain threshold. This threshold is equal to the lower limit of the interval computed by the quartile method. It is defined as MED + k (Q3 - MED), where MED and Q3 are the median and the third quartile, respectively, of the weighted contribution of individuals, and k is a fixed parameter for each province, determined following a review of the highest percentages of contribution to the provincial estimate. (The value of k is 75 for Quebec and Ontario and 25 for all other provinces.)

Having identified the influential values, we must now compute the adjustment factors for their weights. To do so, we determine whether incomes over $99,999 are overrepresented in the sample compared with the distribution of tax-filer incomes from Revenue Canada's T1 file (Canada Revenue Agency). If so, the weight is adjusted so that the distribution of the weighted sample matches that of the tax data (in order to keep weight adjustments to a minimum, this adjustment is not made in cases where the difference between the two distributions is small). The adjustment is made with an adjustment factor whose value is between 0 and 1. A separate adjustment factor is computed for each individual with an influential income, as described below. Note that the calculations are done at the provincial level.

*If a province has more than one influential person, the adjustment factor is computed first for the influential person with the highest income (unweighted). Suppose the person is in income class c. The adjustment will ensure that the estimate of the number of persons in class c or any higher income class is equal to the number of persons in the same income classes based on tax data, subject to certain constraints. First, only the weights of influential persons will be changed by this adjustment. In addition, if the adjustment* *increases the weight, i.e., the adjustment factor is greater than 1, the adjustment will not be applied and the initial weight will be retained. If the tax file shows that there is no one in the income classes, the adjustment will lower the weight to 1. Similarly, the weight will be reduced to 1 if the estimated number of persons in these classes based on non-influential persons only is greater than the number of persons according to the tax file* (Tremblay, 1998).

Calculation of the adjustment factor carries over from one individual to the next. When the adjustment factor has been computed for the influential individual with the highest income, that individual's adjusted weight is considered in computing the adjustment factor for the next individual.

Although the adjustment factor is calculated at the person level, it has to be applied to the weights of the influential individual and all other members of the same household. For households in which no individuals are identified as having extreme income values, the adjustment factor is 1 for all household members.

The method for detecting influential values uses the weight obtained after the calibration step (described in the next section) since this relies on the population income distribution. Therefore, temporary calibration is required before implementing the step for influential values adjustment, using temporarily calibrated weights to identify and process influential observations. Once the adjustment factors have been computed, the calibration step must be redone; the factors are applied to the weights that we had before the temporary calibration (i.e., the cross-sectional weights following the weight share step, $w_{2003,\,\mathrm{shared}}$).

The cross-sectional weight adjusted for influential values is computed as follows:

$$w_{2003,\,infl} = w_{2003,\,\mathrm{shared}} * \boldsymbol{b}_{infl}$$

where $w_{2003,\,infl}$ = cross-sectional weight adjusted for influential values

$w_{2003,\,\mathrm{shared}}$ = cross-sectional weight following the weight share step

$\boldsymbol{b}_{infl}$ = adjustment factor for influential values

When the individuals living with all persons identified as having influential incomes in the current wave or previous waves were counted, the adjustment for influential values was applied to a total of 66 persons for the 2003 reference year. The adjustment factors applied to their weights range from 0.003 to 0.98.

## 5.7 Calibration

The final step in the cross-sectional weighting process is calibration. As noted in section 4.5, this part of the weighting process has undergone many changes since SLID's inception. Most of them were made at the time of the last two historical revisions, which were carried out in 2002 and 2005 following an extensive calibration harmonization project for income and expenditure surveys.

The purpose of calibration is to ensure that the survey estimates for certain subgroups in the sample match known population totals for a given year. In 1993 and subsequent years, growing discrepancies were observed between SLID estimates and estimates from external sources. The Harmonized Calibration of Income Statistics Project was launched in 1999 as a result of these discrepancies, and some differences between the estimates produced by the various income and expenditure surveys. Its aim was to harmonize the calibration strategies of income and expenditure surveys, and thereby improve the comparability of the various estimates produced by surveys and the estimates from external sources (Tremblay, 2005).

For SLID, two major problems with the survey estimates needed to be resolved. One problem had to do with the distribution of the sample by size of households and economic entities. When SLID's pre-calibration estimates were compared with the known population totals, it was discovered that the number of size 2+ households was underestimated by more than 8% (this is referred to as slippage), while the number of size 1 households was underestimated by only about 2% (LaRoche, 2005). It was also found that the numbers of economic entities in the various size categories were underestimated, by far the largest discrepancy being nearly 17% in the size 1 estimate. This imbalance between the estimates of households and economic entities by size resulted in a substantial underestimate of the number of size 1 economic entities living in size 2+ households. When household size and economic entity size were not controlled for in calibration, a significant bias was observed in the corresponding survey estimates.

The second problem with the representativeness of the sample had to do with income. Beginning in the 1990s, growing discrepancies were found between income estimates derived from surveys and estimates from reliable external sources. When the aggregate wages and salaries series from SCF/SLID were compared with the System of National Accounts series, it was evident that between 1980 and 1991, SCF overestimated aggregate wages by some 1.6% to 3.6% compared with the System of National Accounts and that in 1992 and subsequent years, the gap widened to between 3.1% and 6.2% (Lathe, 2005).

Therefore, the distribution of paid employees by wages and salaries class was examined in an effort to determine exactly what was causing this overestimate of aggregate wages and salaries. The results showed that SLID tended to overestimate the number of persons whose wages and salaries fell in the centre of the distribution, and to underestimate the numbers in the tails of the distribution. This tendency was also present for SCF and other surveys (Tremblay, 2005).

In the Harmonized Calibration of Income Statistics Project, many studies were conducted with a view to identifying the best calibration strategy, i.e., the one that would reduce the discrepancies between certain key estimates from different sources while limiting distortion in the weights. Another important criterion in choosing the best strategy was that it had to minimize the negative effects on other survey estimates not controlled by calibration.

A large number of control totals were evaluated in the process of selecting the best ones for the survey. Demographic control totals, such as the number of persons in various age-sex groups, and the number of households and economic entities by size and by type, were evaluated. For income-based control totals, various external sources were considered, such as T1 and T4 files from the Canada Revenue Agency. The T1 file contains data from income-tax and benefit returns filed by individuals, while the T4 file contains data from statements of remuneration paid, which are filed by employers. The many studies conducted showed that the T4 file provided better coverage, especially for low-income persons (Auger and Tremblay, 2005). Consequently, the control totals selected were the number of paid employees per wages and salaries class. The class boundaries are defined in terms of specific percentiles. For SLID, the percentiles used are the $10^{th}$, the $25^{th}$, the $50^{th}$, the $65^{th}$, the $75^{th}$ and the $95^{th}$, $98^{th}$ or $99^{th}$, depending on the number of observations in the top class. Note that the $10^{th}$ percentile class was added for SLID because of the importance of the low-income population for the survey. The choice of the other boundaries was guided primarily by the fact that good representation of higher incomes was important for a better estimate of the aggregate totals.

For SLID, calibration improvements were introduced in two stages. Phase 1 of the calibration harmonization project began when the 2000 reference year was released (May 2002). The changes in the calibration strategy affected only the demographic totals. Phase 2 went into effect when the SLID data for the 2003 reference year were released (May 2005). The most important change in phase 2 is the addition of control totals based on the number of paid employees per wages and salaries class. Also, the number of age groups was reduced to keep the total number of controls used down, and limit the distortion of weights during calibration. The control totals used at the time of each historical revision are shown in Table 4.3 in the section on longitudinal weighting. For more details on the reasoning behind the selection of SLID's current calibration strategy, and on the use of administrative data in the calibration step, see LaRoche (2005) and Latouche (2005).

In the case of cross-sectional calibration, the survey estimates have to be adjusted so that they represent the target population on December 31 of the reference year. Consequently, individuals (cross-sectionally in scope and respondent) are assigned to the various calibration classes on the basis of their characteristics on December 31 of the reference year. While longitudinal calibration has to be performed in two stages, cross-sectional calibration requires only one, since the control totals and the survey variables are available for the reference period. The control totals used for cross-sectional calibration are given in Table 5.2.

For SLID, additional restrictions are applied to the final weights during calibration. First, for the integrated and labour weights, the final weights have to be greater than or equal to 1. Second, for the integrated weight, the other constraint is that all members of a household must have identical weights after calibration (Lemaître and Dufour, 1987). Integration of weights at the household level makes it possible to produce estimates at the household, economic entity and person levels.

**Table 5.2 Control totals used in cross-sectional calibration for the 2003 reference year**

| | Control totals | | |
|---|---|---|---|
| Reference period: December 31, 2003 | Age-sex group | 0 to 6 (both sexes)<br>7 to 17<br>18 to 24<br>25 to 34 | 35 to 44<br>45 to 54<br>55 to 64<br>65 and over |
| | Number of households by size | Size=1and size=2 | |
| | Number of economic entities by size | Size=1and size=2 | |
| | Wages and salaries classes | 7 classes defined by the $10^{th}$, $25^{th}$, $50^{th}$, $65^{th}$, $75^{th}$ and $95^{th}$, $98^{th}$ or $99^{th}$ percentiles | |

On the basis of the control totals shown in Table 5.2 and the theory outlined in section 4.5.1, the cross-sectional weights vector produced by the calibration process is given by

$$\underline{w}_{\text{cal, int}} = \text{diag}\left(\underline{w}_{2003,\text{i}nfl}\right)\underline{g}$$

where $\text{diag}\left(\underline{w}_{2003,\text{i}nfl}\right)$ is an *n* x *n* diagonal matrix with the integrated weights after the adjustment for influential values, $\underline{w}_{2003,\text{i}nfl}$, on the diagonal, and $\underline{g}$ is the *n* x *1* vector of the g-weights produced with the additional constraint that the final weights must be identical for all members of the same household.

For the labour weight, the process is the same except that there is no requirement for all the weights in a household to be the same. The cross-sectional labour weights vector produced by calibration is given by

$$\underline{w}_{\text{cal, lab}} = \text{diag}\left(\underline{w}_{2003,\text{i}nfl}\right)\underline{g}$$

where $\text{diag}\left(\underline{w}_{2003,\text{i}nfl}\right)$ is a*n n* x *n* diagonal matrix with the labour weights after adjustment for influential values, $\underline{w}_{2003,\text{i}nfl}$, on the diagonal, and $\underline{g}$ is the *n* x *1* g-weights vector produced by calibration.

For the 2003 reference year, the integrated cross-sectional weights after adjustment for influential values were multiplied by an average of 1.16 to correct the SLID sample. On the labour side, the cross-sectional labour weights were adjusted by an average of 1.12 to correct the sample.

The two sets of weights produced by calibration are SLID's final cross-sectional weights.[10] Table 5.3 shows the median final cross-sectional integrated and labour weights for the 2003 reference year. Only individuals with non-zero final weights are used to compute the medians. The province shown in the table refers to the province of residence on December 31, 2003.

10. At one time, SLID also produced a set of weights with random noise added for external use (Lévesque and Franklin, 2000). Those weights are no longer produced.

**Table 5.3 Median final cross-sectional weights for the 2003 reference year**

| Province | Median | |
|---|---|---|
| | Integrated | Internal labour |
| | weight | |
| Newfoundland and Labrador | 154 | 160 |
| Prince Edward Island | 65 | 68 |
| Nova Scotia | 162 | 174 |
| New Brunswick | 163 | 176 |
| Quebec | 342 | 373 |
| Ontario | 377 | 407 |
| Manitoba | 189 | 213 |
| Saskatchewan | 162 | 176 |
| Alberta | 453 | 493 |
| British Columbia | 561 | 598 |

## 6 Forthcoming changes and future developments

SLID's longitudinal and cross-sectional weighting steps are regularly revised to improve the quality and reliability of the estimates. The revisions are carried out following studies or assessments, with due consideration for any changes in objectives, processing or dissemination. Various steps in the weighting process are continually being assessed to ensure that the survey is up-to-date with new methodologies and is using the methods best suited to its needs.

SLID and a number of other surveys are facing declining response rates, and the problem is getting worse. New procedures were introduced for SLID's January 2006 collection phase with a view to identifying the various types of respondents more effectively, and increasing the chances of reaching them. For example, if a particular type of household is more likely to be contacted during a specific period of the day, attempts to contact that type of household will be given priority in that period.

While efforts are being made to reduce the effects of non-response by improving collection methods, it is still important to ensure that pertinent, appropriate non-response adjustment methods are used. In particular, various studies will have to be conducted soon to evaluate the current non-response adjustment strategy. They will focus on the following aspects of non-response adjustment: segmentation modelling compared with other methods, such as logistic regression; adjustment at the provincial or regional level instead of the national level; the addition of explanatory variables based on types of non-response (refusal, no contact, etc.); the introduction of geography-based indicator variables to explain non-response; and so on. These studies will help determine whether changes are needed in the current methodology and whether new methods have to be introduced.

Another change that would improve the survey is the development and implementation of an imputation strategy for labour data. This would make it possible to offset non-response on the labour interview by imputing the labour data, and there would no longer be any point in producing the labour weights. Since it is more complex to develop a labour-data imputation strategy, and since such a project would require a significant commitment of methodological and subject-matter resources, the idea has been on hold for many years, but it remains important for the methodology team. Until that labour variables imputation strategy is developed, labour weights will remain necessary and will continue to be produced. However, changes could be made in the definition of a respondent with a labour weight to limit the number of missing variables for users. Various options are currently being considered. In particular, the possibility of redefining respondents as individuals rather than households will be explored. In addition, the possibility of producing labour weights more quickly will have to be studied in order to prepare for the eventuality that a publication containing labour data has to be released early.

SLID has now been running for more than 10 years. Depending on the survey's results, the experience gained from it, feedback from key users and new data needs, a thorough redesign of the survey is in progress and should be implemented in the next few years. Among the features that could be revisited are panel duration, the number of years before a new panel is introduced and the number of overlapping panels.

## 7 Conclusion

SLID's goal of providing longitudinal and cross-sectional data on a number of characteristics associated with labour and income dynamics, and ensuring that those data are representative of the population of Canada's ten provinces is a formidable challenge. The longitudinal and cross-sectional weights needed to meet that challenge are produced through the longitudinal and cross-sectional weighting processes described in this report. There are several steps in those processes: determination of the initial weights, non-response adjustment, adjustments for interprovincial migration (for cross-sectional weights only), the combination of the panels and weight share (for cross-sectional weights only), adjustment for influential values, and calibration. These steps have been improved and optimized over the years, but with the evolution of processing and dissemination strategies, new refinements are needed, in particular to maintain data quality and reliability.

**Acknowledgements**

The author would like to thank Caroline Cauchon, Michel Latouche, Pierre Lavallée and Soumaya Moussa for their valuable advice and comments, which improved the quality of this report.

## Appendix A: List of variables used in the non-response adjustment model

The following dichotomous variables were used to model non-response for wave 5 of Panel 3 and wave 2 of Panel 4.

| | |
|---|---|
| Educ1 | - 0 to 8 years of schooling |
| Educ2 | - 9 to 13 years of schooling (high school) |
| Educ3 | - postsecondary excluding community college, CEGEP and university |
| Educ4 | - community college, CEGEP or university (including Ph.D.) |
| RENTER | - 1=owner, 0=renter |
| EMPLOYE | - employed |
| NON_EMPL | - unemployed |
| NLF | - not in the labour force |
| MARIE | - married or living common-law |
| SEP_DIV | - separated or divorced |
| VEUF | - widowed |
| CELIB | - never married |
| P10 – P59 | - one variable per province |
| SEXE | - sex |
| AGE_0 | - age < 16 |
| AGE_1 | - age >=16 and <19 |
| AGE_2 | - age >=19 and <25 |
| AGE_3 | - age >=25 and <35 |
| AGE_4 | - age >=35 and <45 |
| AGE_5 | - age >=45 and <55 |
| AGE_6 | - age >=55 and <65 |
| AGE_7 | - age >=65 |
| HHSZ1-HHSZ5 | - number of persons in the household (HHSZ5=>5) |
| FAMTYPE1 | - unrelated person |
| FAMTYPE2 | - husband-wife family or lone-parent family with no young children |
| FAMTYPE3 | - husband-wife family or lone-parent family with children aged 0 to 17 |
| FAMTYPE4 | - other family type |
| STUDENTS | - 1=student (full time or part time), 0=not a student |
| CWORK1 | - class of worker = public-sector employee |
| CWORK2 | - class of worker = private-sector employee |
| CWORK3 | - self-employed private-sector worker with or without employees or unpaid family worker |
| MULTIHH | - multiple-family household |
| URBAIN | - lives in an urban area (based on enumeration area) |

**References**

Auger, S. et Tremblay, J. (2005). *Évaluation de la qualité des estimations démographiques et des données fiscales utilisées dans le calage des différentes enquêtes à Statistique Canada.* Recueil de la section des méthodes d'enquête, Assemblée annuelle de la Société Statistique du Canada, juin 2005.

Bankier, M. (2002). *Pondération pour le recensement du Canada 2001.* Modélisation de données d'enquête pour la recherche économique et sociale, Assemblée annuelle de la Société Statistique du Canada, juin 2002

Cotton, C., Bishop, K., Giles, P., Hewer, P. et Saint-Pierre, Y. (1999). *Comparaison des résultats de l'Enquête sur la dynamique du travail et du revenu (EDTR) et de l'Enquête sur les finances des consommateurs (EFC) 1993-1997 : mise à jour*, publication de Statistique Canada, 75F0002MIF – No 99007.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. et Lindeyer, J. (1998). *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, Catalogue 71-526-XPB.

Godbout, S. (2004). *Étude sur l'ajustement pour la non-réponse et la surestimation des ménages individuels*, document interne, Statistique Canada.

Kalton, G. et Kasprzyk, D. (1986). *Le traitement des données d'enquête manquantes*, Techniques d'enquête, vol. 12, no 1, pp. 1-17.

LaRoche, S. (2005). *Survey of Labour and Income Dynamics Calibration Strategy.* Survey Methods Proceedings, Statistical Society of Canada Annual Congress, june 2005.

Lathe, H. (2005). *Enquête sur la dynamique du travail et du revenu : révision historique de 2003,* Série de document de recherche sur le Revenu, Statistique Canada, 75F0002MIF au catalogue — No 009

Latouche, M. (2005). *Leçons à tirer de l'utilisation de données administratives au stade de la pondération.* Recueil de la section des méthodes d'enquête, Assemblée annuelle de la Société Statistique du Canada, juin 2005.

Latouche, M., Dufour, J. et Merkouris, T. (2000). *SLID Cross-Sectional Weighting: Combining Two or More Panels*, Série de document de recherche sur le revenu, Statistique Canada, 75F0002MIE – 00006.

Latouche, M. et Michaud, S. (1997). *Concerns pertaining to weighting of longitudinal surveys*, Proceedings of the Section on Government Statistics and Section on Social Statistics, American Statistical Association.

Lavallée, P. (1994). *Ajout du second panel à l'EDTR: sélection et pondération*, document interne, Statistique Canada.

Lavallée, P. (1995). *Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage de poids*, Techniques d'enquête, vol. 21, no 1, pp. 27-35.

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée de partage des poids*, Édition de l'université de Bruxelles.

Lavigne, M. et Michaud, S. (1998). *Aspects généraux de l'enquête sur la dynamique du travail et du revenu*, document de travail sur la dynamique du revenu et du travail, Statistique Canada, 75F0002M No. 98-05.

Lévesque, I. et Franklin, S. (2000). *Pondérations longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu, année de référence 1997*, document de travail sur la dynamique du revenu et du travail, Statistique Canada, N° 75F0002MPF-00004.

Lemaître, G. et Dufour, J. (1987). *Une méthode intégrée de pondération des personnes et des familles*, Techniques d'enquête, vol. 13, no 2, pp. 211-220.

Merkouris, T. (1999). *Cross-sectional Estimation in Multiple-Panel Household Surveys*, document de travail de la direction de méthodologie, Statistique Canada, HSMD-99-004E.

Naud, J,-F. (2004). *Pondération longitudinale avec panels combinés, 'Enquête sur la dynamique du travail et du revenu, 1996-2002.* Document de travail sur la dynamique du revenu et du travail, Statistique Canada Nº 75F0002MIF2004008

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*, Statistique Canada, Catalogue 71-526.

Statistique Canada (1997). *Répartition du revenu au Canada selon la taille du revenu*, publication de Statistique Canada N 13-207-XPB, pp. 52-59.

Tambay, J. L., Schiopu-Kratina, I., Mayda, J., Stukel, D. et Nadon, S. (1998). *Traitement de la non-réponse du cycle deux de l'enquête sur la santé de la population*, Techniques d'enquête, vol. 24, no 2, pp. 159-169.

Tremblay, J. (1998). *Détection des observations influentes pour l'enquête sur les finances des consommateurs (EFC) et l'enquête sur la dynamique du travail et du revenu (EDTR)*, document interne, Statistique Canada.

Tremblay, J. (2005). *Aperçu de la stratégie de calage harmonisée des statistiques du revenu de Statistique Canada.* Recueil de la section des méthodes d'enquête, Assemblée annuelle de la Société Statistique du Canada, juin 2005.