WORKING DOCUMENT
METHODOLOGY BRANCH

FIRST THREE CYCLES OF THE NATIONAL POPULATION HEALTH SURVEY HOUSEHOLD COMPONENT

DMEM – 2005-017EF

F. Brisebois, J. Dufour, D. Kelly, M. Lavigne, P. Mathieu and S. Tolusso

Household Survey Methods Division
Statistics Canada

September 21, 2005

Statistics Canada Statistique Canada

Canada

FIRST THREE CYCLES OF THE NATIONAL POPULATION HEALTH SURVEY HOUSEHOLD COMPONENT

F. BRISEBOIS, J. DUFOUR, D. KELLY, M. LAVIGNE, P. MATHIEU, S. TOLUSSO

ABSTRACT

The National Population Health Survey (NPHS) is a biennial survey that originated at Statistics Canada in 1994. The main goal of this survey is to measure the health status of Canadians and to promote a better understanding of the factors determining health. To provide a complete portrait of the Canadian population, the NPHS has three components: a household survey, a survey of long-term residents of health care institutions, and a northern survey (which covers households in the territories). This document covers only the first three cycles of the household survey in the ten provinces. During its first three cycles, the survey disseminated not only longitudinal but also cross-sectional results. This paper firstly presents a background of the NPHS and continues by describing its various aspects, including coverage, sampling frame, sample design, content, questionnaire, non-response, weighting, data quality, confidentiality and dissemination.

TROIS PREMIERS CYCLES DE L'ENQUÊTE NATIONALE SUR LA SANTÉ DE LA POPULATION VOLET MÉNAGES

F. BRISEBOIS, J. DUFOUR, D. KELLY, M. LAVIGNE, P. MATHIEU, S. TOLUSSO

RÉSUMÉ

L'Enquête nationale sur la santé de la population (ENSP) est une enquête biennale qui a vu le jour à Statistique Canada en 1994. Cette enquête vise principalement à mesurer l'état de santé des Canadiens et à favoriser une meilleure compréhension des facteurs déterminants pour la santé. D'ailleurs, pour dresser un portrait tout entier de la population canadienne, l'ENSP comporte trois volets: une enquête auprès des ménages, une enquête auprès des résidents à long terme des établissements de soins de santé et une enquête nordique (qui couvre les ménages dans les territoires). Dans ce document, on s'intéresse uniquement aux trois premiers cycles de l'enquête auprès des ménages des dix provinces canadiennes. L'enquête, jusqu'à son troisième cycle d'existence, permet de diffuser des résultats non seulement à l'échelle longitudinale mais également à l'échelle transversale. Ce document présente tout d'abord une mise en contexte de l'ENSP et poursuit en décrivant toutes ses étapes soient la couverture, la base de sondage, le plan d'échantillonnage, le contenu, le questionnaire, la collecte, le traitement, la non-réponse, la pondération, la qualité des données, la confidentialité et la diffusion.

TABLE OF CONTENTS

1. Introduction

The National Population Health Survey (NPHS) is a biennial survey that originated at Statistics Canada in 1994. The main goal of this survey is to measure the health status of Canadians and to promote a better understanding of the determinant health factors. To provide a complete portrait of the Canadian population, the NPHS has three components: a household survey, a survey of long-term residents of health care institutions, and a northern survey (which covers households in the territories). This paper covers only the first three cycles of the household survey in the ten Canadian provinces.

The first NPHS data collection cycle began in 1994. Up to its third cycle of existence, the survey disseminated not only longitudinal but also cross-sectional results. The first contact with households sampled by the NPHS is in person, using computer-assisted interviews. First, basic information is collected on all members of the sampled household, who constitute the *general component* of the survey. Information on behaviour and on self-perception, which constitutes the *health component*, is collected in an interview with just one randomly selected member of the household, who thus becomes a member of the NPHS panel. Together, the panel members constitute the *longitudinal sample*, which will be contacted every two years for eighteen years. The interviews for subsequent cycles are done by telephone using computer-assisted telephone interviews, and in each cycle, the same content is found, supplemented by focus content or a few additional questions.

This report comprises fourteen sections. Section 2 presents the background of the NPHS. Section 3 describes the target population and the sampling frames used. Section 4 describes the sample design of the household component in the first three cycles of the NPHS. Section 5 looks at the questionnaire and its content. Data collection is then discussed in section 6, after which the processing of the data is dealt with in section 7. Section 8 describes the challenges posed by partial and total non-response. This is followed, in section 9, by a brief description of the different products made available to users. Section 10 gives an overview of the methodology for weighting and estimation. Data quality is discussed in section 11. Section 12 describes the confidentiality measures taken for the different products offered to users. Data dissemination is then discussed in section 13. Lastly, section 14 concludes this report by describing the major changes made to the NPHS for future cycles.

2. Background

Prior to 1994, population health surveys were infrequent at Statistics Canada (Catlin and Will, 1992). The last major survey on the health of the population, the Canada Health Survey, had been conducted in 1978. Originally intended to be permanent, that survey ceased operations owing to budget cuts imposed across the federal Public Service in the second quarter of 1978. After that, various surveys filled some of the gaps observed with respect to health, focusing on particular subjects or specific populations. For example, in 1983, Statistics Canada conducted the Canadian Health and Disability Survey, followed by the Health and Activity Limitations Survey in 1986 (Dolson, McClean, Morin and Théberge, 1987). These two surveys were primarily intended to collect specific information on the nature of problems experienced by Canadians with disabilities.

In 1990, Statistics Canada conducted the Canadian Health Promotion Survey (Health and Welfare Canada, 1993) and in 1991, the theme of the General Social Survey was health. However, the scope of these two surveys was quite limited (Yeo, 1999). None of these surveys provided a complete picture of the health status of the population and the many factors that affect health.

In the fall of 1991, the National Health Information Council recommended that a permanent national survey be conducted on the health of the population. This recommendation was made in response to economic and fiscal pressures on the health care system and the need for information to improve the health status of the Canadian population. Following on this recommendation, Statistics Canada received the mandate and funding to develop a longitudinal health survey that was to be flexible and capable of producing valid, reliable and timely results. The survey also had to be able to adapt easily to changing needs, policies and interests. In 1994, the National Population Health Survey (NPHS) was in the field for the first time.

The NPHS has numerous objectives:

i) to aid in the development of government policies designed to improve health status;
ii) to provide data for analytic studies leading to a better understanding of the determinants of health;
iii) to collect data on the economic, social, demographic, occupational and environmental correlates of health status;
iv) to increase the understanding of the relationship between health status and health care utilisation;
v) to provide data on a constant sample that will reflect the dynamics of health and illness, and to produce periodic cross-sectional estimates;
vi) to provide the provinces and territories and other clients with a collection mechanism that will permit supplementation of content or sample;
vii) to allow the possibility of linking data with administrative sources.

The NPHS produced both cross-sectional and longitudinal estimates during the first three cycles of its existence. Starting with the fourth cycle, the cross-sectional component was taken over by the Canadian Community Health Survey (Béland, Bailie, Catlin and Singh, 2000); the NPHS then became a strictly longitudinal survey.

3. Target Population and Sampling Frames

The target population of the NPHS household component includes all persons living in private dwellings who reside in the ten provinces. Excluded from the scope of the survey are persons living on Indian reserves or Crown lands, institutional residents, full-time members of the Canadian Forces and residents of some remote areas.

The choice of a frame for sample selection depends on a number of factors, but the frame must first and foremost correspond as closely as possible to the survey's target population. Moreover, the creation, use, updating and verification of the sampling frame must adhere to the survey's operational and financial constraints. To meet all these criteria, a decision was made to use three different sampling frames.

3.1 Sampling Frame of the Labour Force Survey

First, the area frame developed for the Labour Force Survey (LFS) (see Singh, Drew, Gambino and Mayda, 1990 and Singh, Gambino and Laniel, 1994) was used as the main frame for selecting the NPHS sample for every province except Quebec. The area frame developed for the LFS offered a number of definite advantages for selecting the NPHS sample. These included an infrastructure already in place for making updates to take account of new dwellings, demolished dwellings and out-of-scope units, as well as the entire process for evaluating the coverage of the frame. Also, since several other household surveys also use this area frame, it is easier to control sampling overlaps. With this frame, it is possible to select either a sample of new dwellings or a sample of rotated out dwelling from the LFS. Rotated out dwellings are dwellings that spent six months in the LFS sample and ended their participation in that survey. The LFS selects dwellings from the area frame according to a multistage stratified clustered sample design. Given the longitudinal nature of the NPHS, a decision was made to select a sample of new dwellings so as to lighten the response burden.

3.2 Sampling Frame of the Enquête sociale et de santé

In Quebec, the NPHS sample was collected from a second sampling frame, consisting of the dwellings that participated in the Enquête sociale et de santé (ESS), conducted by Santé Québec in 1992-1993 (Courtemanche and Tarte, 1987). This option offered reciprocal advantages: Santé Québec obtained longitudinal coverage for households that agreed to the sharing of NPHS data, and the NPHS could use ESS data to improve the representativeness of its sample without having to reject households on selection (the "rejective" method will be discussed in subsection 4.1.2). The ESS contained 16,010 dwellings, selected according to a two-stage design.

3.3 Random Digit Dialling Frame

The third sampling frame used was one based on random digit dialling (RDD). With this frame, the NPHS permanently has a structure for collecting data by telephone, which can be used to respond quickly to regional and/or provincial requests to purchase sampling units and/or specific content (a practice known as "buy-in), which may differ from one cycle of the survey to another. The RDD frame may be used as a supplement to the LFS or ESS frame, or it may serve as the main frame for meeting special requests.

4. Sample Design

During its first three cycles, the NPHS had to adjust its original sample design primarily to respond more effectively to various cross-sectional needs. These included better representing the cross-sectional population (by adding immigrants and children born in 1995 or thereafter), meeting the changing needs of the provinces (buy-in samples), and meeting the requirements for integration with the National Longitudinal Survey of Children and Youth (NLSCY). Both the development of the original sample design and these many changes posed a number of methodological challenges in terms of multiple sampling frames, the sample design itself, collection and weighting. For more information on the NLSCY, refer to Statistics Canada (2003a).

4.1 Cycle 1

The NPHS sample design was mainly dictated by four initial parameters. The sample sizes planned at the national and provincial scales were the first parameter considered. The second parameter consisted of selecting only one person per household for the health component. The third consisted of using multiple sampling frames. And lastly, the NPHS was to be integrated with the NLSCY.

4.1.1 Size and Sample allocation

The budget allocated to the NPHS called for a sample of approximately 19,600 respondent households. It was also agreed that a minimum of 1,200 households per province were needed in order to obtain reliable estimates for predefined age-sex groups (0-11, 12-24, 25-44, 45-64, 65+). This requirement was taken into account when determining sample sizes for each province using the Kish distribution, well known for producing accurate estimates at both the national and the provincial scale. This is an allocation proportional to $\sqrt{0.804w_h^2 + 1/12^2}$ where w_h represents the proportion of households in the province according to the 1991 Census and h=1,…,10 corresponds to the ten provinces.

The NPHS offered the provinces the opportunity to increase their sample size by buying supplementary sampled units (this type of sample is discussed below). Each province was stratified according to demographic and socioeconomic variables in order to increase the precision of the estimates. The primary goal was to obtain, in each stratum, a sample size proportional to its population in the 1991 Census.

4.1.2 Sample Design Based on the LFS (All Provinces Except Quebec)

The LFS target population includes all persons living in private dwellings in Canada except residents of Indian reserves, institutional residents, full-time members of the Canadian Forces and residents of some remote areas. The LFS survey design is a multistage, stratified clustered design. In each province, three groups are formed. The first group consists of major urban centres. Clusters containing 150 to 250 households are created and stratified according to geographic and socioeconomic characteristics. A few of these sectors include separate apartment strata and strata consisting of census enumeration areas (EAs) identified as having high average income households. Six (or sometimes twelve or eighteen) clusters or apartment buildings are selected from each stratum using a randomized probability-proportional-to-size (PPS) sampling scheme, where size is the number of households.

Urban towns and rural areas form the other two groups in each province. They, too, are stratified by socioeconomic and geographic characteristics. In both these groups, the clusters generally correspond to intersections of LFS-defined employment insurance regions and economic regions. In most strata, six clusters (usually census EAs) are selected using a PPS scheme. In a few areas with low population density, a three-stage design is used. The first stage consists of selecting two or three primary sampling units (PSUs), usually groups of EAs. Then each PSU is divided into clusters. Finally, a sample of six clusters is selected. Six clusters are used in order to allow a one-sixth rotation of the sample every month for the LFS (called a *rotation group*). At each stage, selection is done with PPS. For more information, see Statistics Canada (1998c).

The LFS sample design is set up to yield about 60,000 households every month. Surveys that use the LFS sample design and require smaller sample sizes usually "reserve" from one to six rotation groups per province, with a rotation group being one-sixth of the total sample. To maintain the sample at desired levels, sample stabilisation is used. In other words, rotation groups are subsampled, as when two rotation groups are reserved but the sample size needed represents only 1.5 rotation groups.

Requirements specific to the NPHS led to two modifications to this sampling strategy. The number of "reserved" rotation groups needed was specified at the stratum level rather than the provincial level in order to meet the specific sub-provincial sample size requirements for cross-sectional purposes in the first cycle. It was also required that the number of clusters selected per stratum be a multiple of four (with a minimum of eight) for variance estimation and seasonal representativeness (allowing each stratum to have two or more independent samples of four clusters each7—one per collection period). Each NPHS collection cycle is spread over a period of just over twelve months, which is divided into four periods (see subsection 6.1 for further details). As the NPHS usually requested only between two and six clusters per LFS stratum, similar LFS strata were grouped to form larger NPHS strata with the required number of sample rotation groups. Once the strata were grouped, their rotation groups were also grouped to form replicated samples.

As a result of these changes, the NPHS sample of clusters (rotation groups) can be considered as a stratified replicated sample, where strata are groups of LFS strata and replicates are typically independent, identically distributed samples of four clusters each. There were exceptions, but they are not expected to have a significant impact on survey results.

The number of persons selected in a household and the method of selection of individuals are two other challenges facing the NPHS. A number of health surveys collect the desired information from only one person in the household; other surveys opt to interview all members of the household. The first three cycles of the NPHS are midway between these two options. Socioeconomic data and some health-related information are collected for all members of the household, and this information constitutes the *general component* of the survey. Detailed information on health is collected for only one, randomly-selected member of the household (called the *selected person*), and the information thus collected constitutes the *health component.*

Clearly, such an approach has a financial drawback, since it is necessary to contact enough households to obtain the required number of selected persons. Furthermore, there is a possibility of including an excessive number of members of small households, since the probability of being selected is inversely proportional to the number of persons in the household. Such an approach under-represents persons living in large households (typically parents with children) and over-represents small households (more often than not, single or elderly persons).

To overcome this drawback, a "rejective" method was developed to ensure that large households were adequately represented in the sample. This method involves the rejection of a portion of the households with no one under 25 years of age. The targeted sample sizes were increased to compensate for households rejected because of the method used (Tambay and Mohl, 1995).

Following the rejective approach, part of the sample was subjected to a selection screening to exclude households with no one under 25 years of age from the survey. (It should be noted that the method was applied to all provinces except Quebec, where the sample was obtained from a recent survey and used the information collected to improve the selection of households.) A provincial compensation factor of $1/(1-P_d)$, where P_d represents the percentage of households not expected to be rejected, was applied to the size of the samples. P_d was generally determined for each province and was then applied at the scale of the strata within the provinces. This means that the sample size was increased for strata including a lower percentage of households with no one under age 25 and decreased for strata including a higher percentage of households with no one under age 25. Provincial sample sizes remained the same, but the proportional distribution within the strata was slightly modified.

For cost and operational reasons, the percentage of preselected households (households to which the rejective method was applied) ranged between 25% and 30% in Ontario. In the urban centres of the other provinces, the percentage ranged between 36.5% and 40%, while in rural areas, it ranged between 25% and 30%. The percentages were lower in rural areas, owing to the cost incurred in contacting households. The percentage was also lower in Ontario because there, the rejective method was not applied for the buy-in of additional sampled units. Also, rather than applying the rejective method to apartment strata, where households are generally smaller, the sample sizes were simply reduced.

4.1.3 Sample Design Based on the ESS (for Quebec only)

The 1992-1993 ESS sample includes 16,010 dwellings. This sample uses a two-stage sample design similar to that of the LFS. Quebec was divided into geographic regions by crossing fifteen health regions with four urban density classes (Montreal census metropolitan area (CMA), regional capitals, small urban areas, rural areas). The clusters were formed in each region according to socioeconomic characteristics and were selected with PPS. The clusters thus selected were listed. Random samples of twenty or thirty dwellings per cluster were selected except in major cities, where ten dwellings per cluster were selected.

Santé Québec provided the NPHS with information enabling it to group sampled households into four categories: one-person households, households with children (under 12 years of age), other households with youths (under 25 years of age) and the remaining households (more than one member but no children or youths). For households not responding to the ESS, the NPHS randomly imputed a household category based on the distribution of respondent households in the same cluster. The size of the NPHS sample for Quebec was distributed among the four urban density classes. The distribution was proportional to $\sqrt{2w_h^2 + 1/4^2}$, where w_h represents the distribution of the population according to the four urban density classes h (h = 1, 2, 3, 4). With such a procedure, Montreal does not have too large a sample in relation to the rest of the province. A sufficiently large number of ESS households were selected to be able to provide the required number of households with children, which are the most under-represented household category. The number of households selected in each class was 50% greater than the required

sample size in the class. A subsample of the ESS was created by eliminating two-thirds of one-person households, half of households with no children or youths and one-sixth of households with youths but no children. This subsample was proportional to the population for the four household categories, based on the selected person.

The subsample for Quebec had to be altered with a view to seasonal representation, variance estimation and integration with the NLSCY (see subsection 4.1.5). To form replicates, the ESS strata were grouped. The four collection periods were covered by one group of clusters in each replicate. The clusters themselves covered only one collection period, except for clusters in rural areas and towns since the size was too large.

The sample of households with children was divided into two groups, "Adult" and "Child," according to a ratio of 3 to 2. The "Child" households selected for period 1 or 2 were interviewed in period 3 or 4. For periods 3 and 4, the sample of childless households was also divided according to a ratio of 2 to 3, between an "Adult" sample and a "Child" sample. This means that children who were born or who joined the household between the time of the ESS and the time of the NPHS had a chance of being selected. If, in a "Child" household, no child was present, a person 12 years of age or older was selected, since the rejective method was not used in Quebec.

4.1.4 Purchase of Supplemental Samples

Four provinces (New Brunswick, Ontario, Manitoba and British Columbia) indicated their intention to augment their sample size in order to better represent specific sub-populations. In each case, the sample increase was allocated to health regions. These are sub-provincial geographic regions that the provinces use for administrative purposes. Households selected for supplemental (buy-in) samples are not part of the longitudinal sample, which means that these samples will not be followed in future cycles. The data collected for supplemental samples will be used solely for cross-sectional studies.

In Ontario, supplemental sample was added in each health region (totalling 2,183 households) in order to obtain reliable estimates for three pre-defined age-sex groups per region (see Table 4.2). Manitoba increased its sample to obtain 450 households for the Winnipeg region and 225 households in the other health regions. In both Manitoba and Ontario, the northern regions were treated as a single region so as to keep buy-in samples down and also because of their sparse population. New Brunswick increased the size of its sample in three health regions (in total, 180 households). British Columbia bought 850 additional units (households) for the Prince George region.

Selection of the buy-in samples for Ontario, Manitoba and New Brunswick was based on the LFS frame, using the same sample design as described in 4.1.2. The extra interviews necessitated by the buy-in sample in the Prince George region in British Columbia were too numerous to be conducted by that region's team of LFS interviewers. Therefore RDD was used to select the buy-in sample.

Statistics Canada's RDD frame was used to select most of the buy-in sample for British Columbia. The sampling based on this frame uses the Elimination of Non-Working Banks (ENWB) method. A bank is defined by the regional code followed by the first five digits of the telephone number. A bank is eligible if it includes at least one residential telephone line. The strata are formed by grouping eligible banks together. A number between 00 and 99 is then selected. The combination of the 8-digit bank and the two randomly selected digits yields a

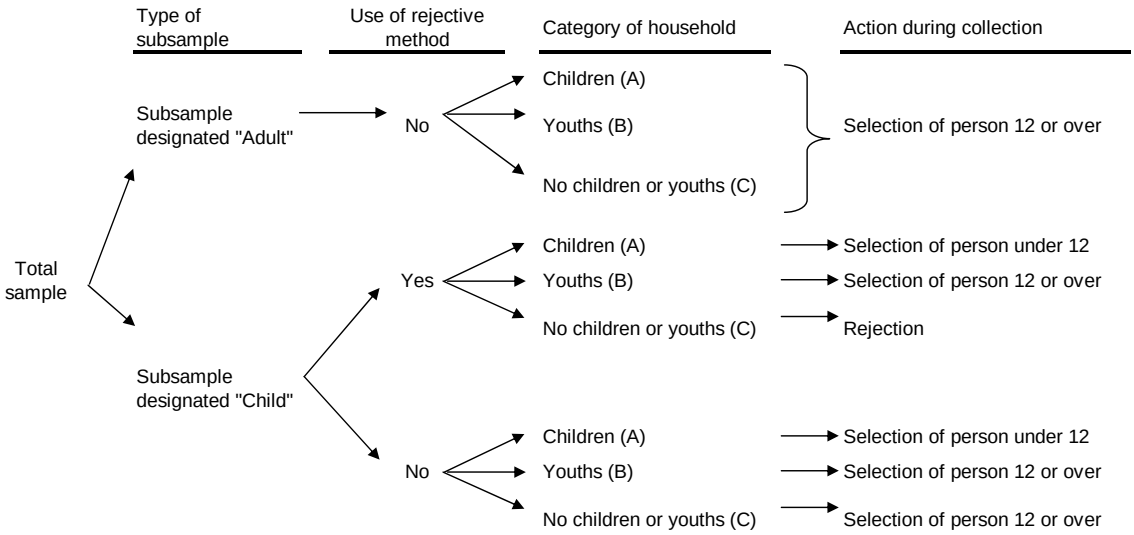
complete, 10-digit telephone number. A historical file ensures that the telephone numbers were not used recently by other RDD surveys. This procedure is repeated until the desired number of telephone numbers in each stratum has been reached. Since it often happens that the telephone number generated is not in service, additional telephone numbers are generated to obtain the required sizes. In general, nearly half the telephone numbers generated belong to households. For more details on this method, see Norris and Paton (1991).

With the buy-in of additional units in the four provinces, it was necessary to obtain 2,840 more respondents.

4.1.5 Important Features of the Sample Design

The Cycle 1 sample design had to be modified somewhat, in consideration of two important features of it: i) the integration of the NLSCY with the NPHS for the data collection, and ii) the under-representation of persons living in large households, owing to the fact that only one person per household is selected for an interview. Figure 4.1 summarizes the strategy adopted. The following subsections provide further details explaining these two features, shedding light on the information presented in Figure 4.1.

Figure 4.1: Distribution of Sample



4.1.5.1 Integration with NLSCY

When the NPHS was being conducted, another survey designed to collect similar characteristics of children was launched. In a joint effort to reduce the response burden on individuals and reduce collection costs, the NPHS and the NLSCY were integrated into a single survey. This integration involved a joint data collection for children, the use of a single collection application and, for the NLSCY, a reduction of the sample size.

The NLSCY is a longitudinal survey of household with a sample size representing 25,000 children under 12 years of age. The sample was selected from households with children which had recently participated in the LFS. Data collection for the first cycle of the survey took place from December 1994 to February 1995; subsequently the survey was conducted every two years. The NPHS could provide a sample of 4,000 to 5,000 children, and consequently the integration enabled the NLSCY to reduce its sample size. In households for which the child was the selected

person, the detailed questionnaire for children was requested of all children in the household, to a maximum of four. After NPHS data collection, the NLSCY processed the detailed data. The NLSCY used data collected from all children, whereas the NPHS used only data collected from children identified as persons selected for the NPHS.

The NPHS timetable required that the survey be in the field well before the NLSCY, which meant that NPHS children could not be selected before the third collection period. This distorted the seasonal representativeness of the sample of selected children and reduced the sample size. To increase the sample yield of children without affecting the seasonal representativeness of other household members in the last two collection periods, part of the NPHS sample for the first two periods was reallocated to collection periods 3 and 4.

To carry out this redesign of the sample, the sample was partitioned in two and then different sampling rules were applied. Part of the sample was therefore designated as the "Adult" subsample and the other part as the "Child" subsample (see column entitled "Type of sub-sample" in Figure 4.1). Unlike what these designations might suggest, three types of households might be included in each subsample: A – Households with children (under age 12), B – households with youths (aged 12 to 25) but no children, and C – households with no youths or children.

In the "Adult" subsample type, only persons aged 12 and over could be selected, regardless of the household type (A, B or C). For the "Child" subsample type, when an interviewed household was of type A (with children), a person aged 12 or under was selected. The selection of persons for the other household types is discussed in the following subsection, since it better suited the needs of the under-representation of large households than of integration with the NLSCY.

Thus the NLSCY could be integrated with the NPHS merely by shifting the collection for the "Child" subsample, planned for periods 1 and 2, to periods 3 and 4. However, given the limited number of interviewers in Prince Edward Island, it was impossible to shift households to another collection period. As a result, the "Child" subsample for that province appears only in periods 3 and 4, which causes seasonal differences in sample sizes for those 12 years of age and over.

4.1.5.2 Under-representation of Persons Living in Large Households

The fact that the sample is obtained by selecting a single person per household generates an under-representation in the sample of persons living in large households (typically couples with children). The reverse is also true, meaning that there is an over-representation of small households (typically single persons, elderly persons or childless couples). To circumvent this problem of representation in the sample, a rejective method was adopted.

The rejective method was used on only a portion of the "Child" subsample (see column entitled "Use of rejective method" in Figure 4.1); it consisted of rejecting households from the sample when they did not include youths or children (hence, type C households). For operational reasons, the portion of the subsample subject to rejection was limited to 25-30% in Ontario, 37.5-40% in other urban areas, and 25-30% in rural areas. For apartment strata, dwellings identified as high income and remote areas, the rejective method was not applied.

It should lastly be noted that for other types of households not discussed thus far, a person aged 12 or over was selected, as shown in the column entitled "Action during collection" in Figure 4.1.

4.1.6 Longitudinal Sample

The longitudinal sample, also called the *longitudinal panel* or simply the *panel*, is made up of 17,276 persons selected in Cycle 1 (including children interviewed by the NLSCY) who at least completed the general component of the Cycle 1 questionnaire (see Table 4.1 for the provincial distribution). It was this longitudinal sample that was followed in cycles 2 and 3, and it will be followed in subsequent cycles of the NPHS. It should be noted that the supplemental samples added for cross-sectional purposes (section 4.1.4) are not included in the longitudinal sample.

The longitudinal sample is not renewed over time. No panel member was, or will be, classified as out-of-scope. The longitudinal sample size will therefore remain 17,276 for all cycles.

Table 4.1: Panel sample sizes

Province	Panel sample size
Newfoundland	1,082
Prince Edward Island	1,037
Nova Scotia	1,085
New Brunswick	1,125
Quebec	3,000
Ontario	4,307
Manitoba	1,205
Saskatchewan	1,168
Alberta	1,544
British Columbia	1,723
Canada	**17,276**

4.2 Cycle 2

Since the NPHS is primarily intended as a longitudinal survey, Cycle 2 mainly consisted of collecting data from the panel of 17,276 persons, as defined in section 4.1.6. However, as in Cycle 1, some provinces bought additional sampled units to be able to produce cross-sectional estimates that were reliable at the subprovincial level.

4.2.1 Longitudinal Sample

Regarding the longitudinal part of the NPHS, no sampling activity took place for Cycle 2. The panel of 17,276 persons defined in Cycle 1 was merely contacted again to collect longitudinal data.

4.2.2 Cross-sectional Sample

As noted in section 4.2.1, no addition was made to the longitudinal sample created in Cycle 1. The panel members are therefore traced to conduct an interview. The general component is first administered to the panel member and to all persons residing with him/her. The entire group of individuals interviewed for the general component thus yield a cross-sectional portrait of the 1996-1997 population.

The health component is then administered only to the panel member. Consequently, the population covered by the Cycle 2 health component is aged 2 and over and includes no immigrant who entered Canada during the past two years. From a longitudinal standpoint, the fact that these subpopulations are not covered poses no problem, since the purpose of the panel is to

be representative of the total population for 1994-1995. However, cross-sectional representativeness for the year 1996-1997 is affected. In this way, the panel alone can only provide a cross-sectional sample aged 2 and over, excluding the population of immigrants who entered Canada since 1994-1995. With the exception of the three provinces that bought a supplemental sample for Cycle 2 (see 4.2.3), this undercoverage of the population is therefore a limitation for the analysis of cross-sectional data from the health component for 1996-1997.

4.2.3 Cycle 2 Buy-in Samples

For Cycle 2, the provinces of Ontario, Manitoba and Alberta bought additional units for cross-sectional purposes. These supplemental samples were initiated and funded by the provincial governments to obtain cross-sectional estimates that were reliable at the subprovincial (health region) level. Because the buy-in was substantial, supplemental samples are dealt with separately from the core sample.

Just as in the case of the core sample, the general component was completed for all members of the households contacted, then the health component was administered to a member aged 12 or over, selected randomly. Also, where possible, in Alberta and Manitoba only, a child aged 0 to 11 was also selected to respond to the health component. Once the Cycle 2 collection was completed, the data collected from the supplemental samples was combined with data from the core sample, thus creating large cross-sectional files.

The sampling of these additional units was done using the RDD frame, similarly to Cycle 1 (see subsection 4.1.4). For Ontario, a sample size of 1,200 respondents (agreeing to share their data) was considered necessary for each of the 23 health regions. Toronto and Ottawa were the exceptions, requiring respectively 3,000 and 2,000 respondents.

In Manitoba, the size of the supplemental sample was determined in such a way that the core and supplemental samples combined would meet a specified level of precision. In the province's two northern regions, a total of at least 600 respondents for the health component was considered necessary (with the core sample already providing approximately 30 respondents), while in Winnipeg, just over 1,400 additional respondents were required. A total sample of 1,000 was needed in the other eight regions.

Lastly, for Alberta, several factors were taken into consideration in determining the sample sizes required for the province's seventeen health regions. First, the precision required was defined at the level of each health region. The sample size had to yield reliable estimates of proportions of roughly 15% for health regions according to selected age-sex groups. Design effects and the anticipated number of respondents for each population group in the health regions were used to determine the number of households needed. This number depended on the health region as well the regions' requirements, and it ranged between 230 and 2,450 households.

In all, supplemental samples yielded an additional 66,123 respondents (see Table 4.2). Further details concerning supplemental samples are included in the documentation on the public use microdata file for Cycle 2 (Statistics Canada, 1998a).

4.3 Cycle 3

Just as in cycle 2, the longitudinal sample determined in Cycle 1 is again contacted for Cycle 3. For cross-sectional purposes, a top-up sample is added so as to be able to represent sub-populations not covered by the panel for the health component.

4.3.1 Longitudinal Sample

Only panel members are re-contacted for the longitudinal sample. Once traced, the panel member is administered the questionnaires for the general and health components. All persons residing with the panel member at the time of the Cycle 3 interview are interviewed for the general component, solely for cross-sectional purposes.

4.3.2 Cross-sectional Sample

Since all persons residing with the panel member are interviewed for the general component, the cross-sectional sample for that component adequately covers the Canadian population for 1998-1999.

However, because the health component was administered only to panel members, the 1998-1999 cross-sectional population could not be fully covered. Since at the time of Cycle 3, the members of the panel were now 4 years of age and over, the population of 0-3-year-olds was not at all represented in this cross-sectional sample. Similarly, persons who had immigrated to Canada since 1994-1995 were also unrepresented cross-sectionally for the health component. For this reason, a top-up sample was added to the core sample to better serve Cycle 3 cross-sectional needs.

In fact, two top-up samples were selected for Cycle 3. The first consisted of individuals who were not part of the Canadian population in 1994-1995, who could be called a part of the population that was “initially absent.” The second sample was selected to compensate for attrition in the longitudinal sample since Cycle 1. These top-up samples were combined with the core sample to create the general and health cross-sectional samples.

The “initially absent” sample was constructed using children born in 1995 or thereafter (*infants*), as well as immigrants who had arrived in Canada since the start of 1995. January 1, 1995 was used as the cut-off date for these two groups. Four LFS rotated out groups (i.e., those who had completed their six months of participation in this survey) were used as the sampling frame for this population; a single rotation group was used per collection period. When a rotation group was at the end of its participation in the LFS (i.e., when it was in its sixth month), a questionnaire additional to the LFS questionnaire was completed. The country of birth was then collected for each household member. In this way, recent immigrants could be identified. LFS household composition was also used to identify infants.

In all, 422 immigrant households were interviewed by the NPHS, three to four months after their last participation in the LFS. Households with infants were selected from only two of the rotation groups and were interviewed in periods 2 and 3 owing to an operational conflict with the NLSCY. There were 758 households with infants. This number is half of the total number available from the two rotation groups. In collection period 3, the NLSCY needed a small sample of children under 1 year of age, and interviews were conducted by the NLSCY rather than the NPHS. A special stabilization weight was developed for this particular situation and for

households with more than one child.

Just as for the core sample, all members of the households contacted completed the general component. The selected person in each household then continued the interview, answering the questionnaire for the health component. Since the list of members of the household to be contacted was known in advance following the LFS interview, the selected person who had to answer the health questionnaire was chosen even before the files were sent into the field for data collection. Tracing was done if the selected person had moved within three or four months following the LFS interview. It should be noted that for households with infants, the person selected for the health component had to be an infant. Lastly, interviews with immigrants were conducted in person, while those focusing on infants were conducted by telephone to reduce costs.

The second top-up sample was created to compensate for the loss due to attrition since Cycle 1. The households in this sample were selected from the units of the original sample that had not participated in 1994-1995. Since only dwelling addresses were available and household composition was not, it was necessary to visit these dwellings in person. In all, 2,598 households were re-contacted. Because of a problem with the computer application during collection, no persons aged 0 to 11 in this top-up sample were selected to complete the health component. Weighting adjustments were made in order to correct this situation.

4.4 Summary of Cross-sectional Samples

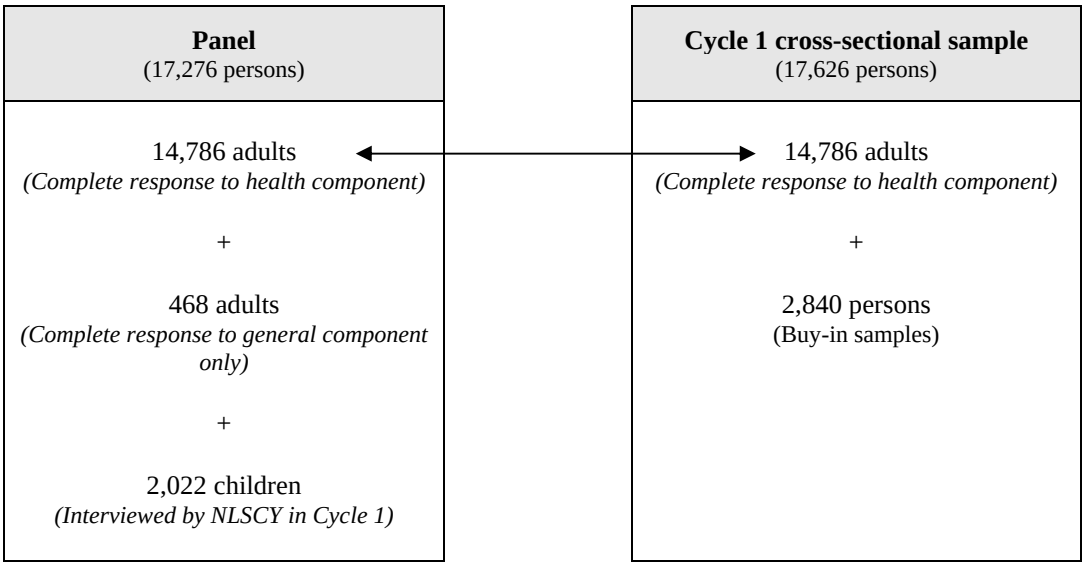
As may be seen above, the main problem in sampling for the NPHS is to provide representative cross-sectional samples based on a fixed longitudinal survey. For the general component, this cross-sectional representativeness does not really pose a problem, owing to the fact that all household members residing with the longitudinal member are interviewed. Since the composition of these households varies from one cycle to another, it is therefore possible to obtain a sample that covers the population adequately. As to the health component, since the longitudinal sample is never renewed, it alone cannot be counted on to represent completely the population in cycles 2 and 3. As one way to fill the gaps, top-up samples were added. Buy-in samples were also added to meet the requirements of the provinces.

Table 4.2 shows the number of respondents obtained after collection and present in the data files disseminated by the survey. This table also shows the distribution of these respondents by their origin (i.e., whether they came from the core sample or a buy-in or top-up sample). Figure 4.2 shows the relationship between the panel and the Cycle 1 cross-sectional sample.

Table 4.2 Cross-sectional Sample Sizes (Number of Respondents) for Different Cycles: Core, Buy-in and Top-up Samples

Cycle/ province	Cross-sectional sample sizes							
	General component:				Health component:			
	Core	*Buy-in*	*Top-up*	*Total*	*Core*	*Buy-in*	*Top-up*	*Total*
CYCLE 1								
Nfld.	3,511	-	-	3,511	918	-	-	918
P.E.I.	3,106	-	-	3,106	899	-	-	899
N.S.	3,071	-	-	3,071	911	-	-	911
N.B.	3,188	419	-	3,607	974	137	-	1 111
Que.	8,461	-	-	8,461	2,581	-	-	2,581
Ont.	12,056	5,165	-	17,221	3,664	1,523	-	5,187
Man.	3,384	1,360	-	4,744	1,025	395	-	1,420
Sask.	3,161	-	-	3,161	1,005	-	-	1,005
Alb.	4,487	-	-	4,487	1,310	-	-	1,310
B.C.	4,696	2,374	-	7,070	1,499	785	-	2,284
Canada	49,121	9,318	-	58,439	14,786	2,840	-	17,626
CYCLE 2								
Nfld.	3,017	-	-	3,017	963	-	-	963
P.E.I.	2,752	-	-	2,752	918	-	-	918
N.S.	2,775	-	-	2,775	986	-	-	986
N.B.	2,888	-	-	2,888	1,032	-	-	1,032
Que.	7,838	-	-	7,838	2,788	-	-	2,788
Ont.	10,899	99,946	-	110,845	3,867	35,527	-	39,394
Man.	3,045	29,354	-	32,399	1,101	13,727	-	14,828
Sask.	2,785	-	-	2,785	1,047	-	-	1,047
Alb.	4,164	36,638	-	40,802	1,436	16,869	-	18,305
B.C.	4,276	-	-	4,276	1,543	-	-	1,543
Canada	44,439	165,938	-	210,377	15,681	66,123	-	81,804
CYCLE 3								
Nfld.	2,560	-	311	2,871	864	-	99	963
P.E.I.	2,457	-	298	2,755	836	-	96	932
N.S.	2,529	-	448	2,977	912	-	159	1,071
N.B.	2,615	-	344	2,959	950	-	123	1,073
Que.	7,091	-	1,214	8,305	2,542	-	404	2,946
Ont.	10,232	-	3,294	13,526	3,632	-	1,059	4,691
Man.	2,878	-	481	3,359	1,019	-	165	1,184
Sask.	2,600	-	492	3,092	970	-	161	1,131
Alb.	4,028	-	513	4,541	1,405	-	163	1,568
B.C.	3,792	-	869	4,661	1,390	-	295	1,685
Canada	40,782	-	8,264	49,046	14,520	-	2,724	17,244

Figure 4.2: Relationship between Panel and Cycle 1 Cross-sectional Sample (Health Component)



5. Questionnaire and Content

NPHS questionnaires are constructed using a block approach. Four different blocks are listed: *input-output*, *demographic*, *general component* and *health component*. These blocks contain different types of content: core, focus, buy-in and supplemental. There are numerous questions dealing with these types of content, including two especially sensitive questions: the question on data sharing and the question on permission for data linkage. These two questions allow the creation of different survey products and will be described in greater detail in subsection 5.3. Lastly, to ensure that the collection mechanism functions properly, several tests are carried out and are described in subsection 5.4.

5.1 Blocks of the Questionnaire

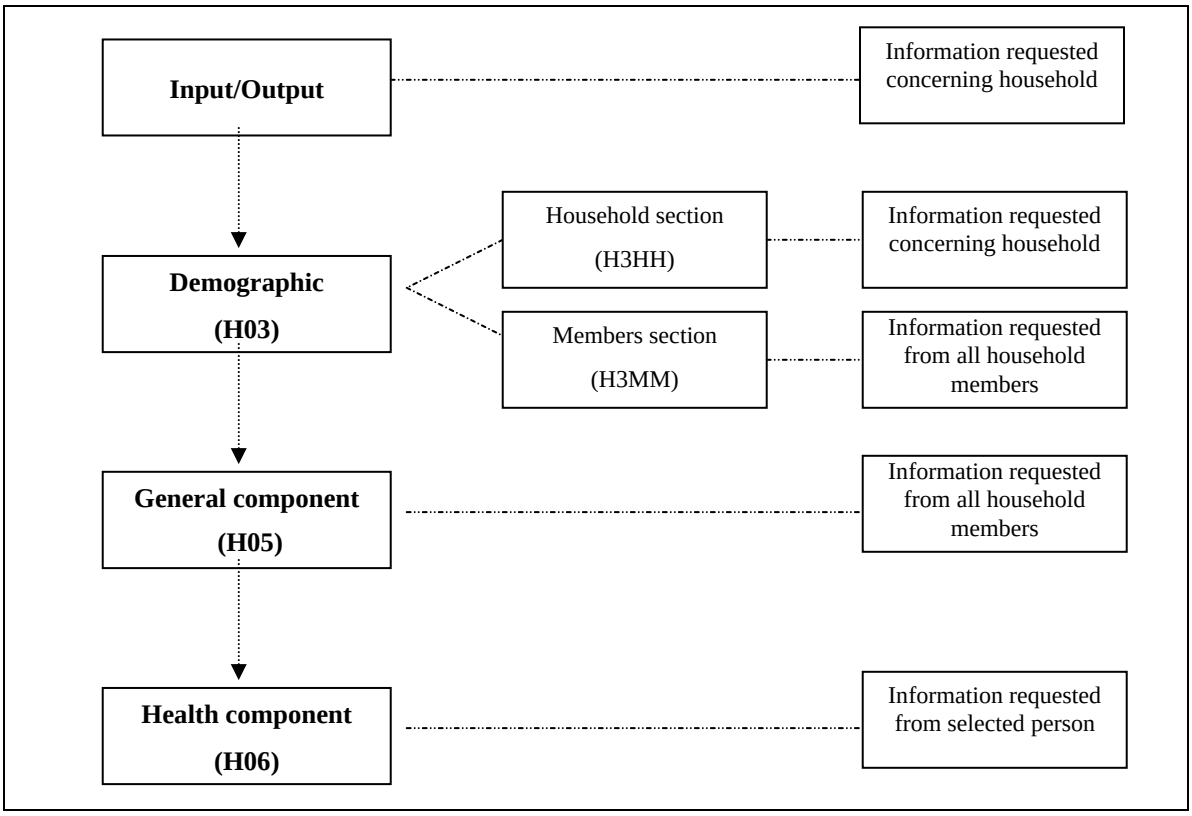
The NPHS questionnaires for cycles 1, 2 and 3 include four blocks (see Figure 5.1). The first block is called *input-output* and includes two sections: input and output. The input questions ensure that the household is indeed eligible for selection. For the longitudinal respondent, this section of the block serves to verify whether he/she is still living in the same place (for cycles 2 and 3). If the longitudinal respondent has moved, any new information for locating him/her is collected here. In the output section, the provincial health insurance number is collected, along with permission to link the data collected with information held by provincial health ministries and consent for the information to be shared with our partners. Also, at the end of the interview, each longitudinal respondent is asked for information regarding two contact persons (usually a friend, neighbour or relative). This information will be used, if necessary, to do follow-up interviews in future cycles. The NPHS contacts these persons only if the longitudinal respondent has moved and the interviewer is unable to obtain a new address from the input section of the block.

Next comes the *demographic* block *(H03)*, which contains the *household* section *(H3HH)*. Here information is collected on dwelling type, owner/renter status, number of bedrooms, address and telephone number. The other section of the demographic block is called the *members* section *(H3MM)*, which contains information on household composition. This section contains the list of all household members as well as the relationships among them. It also contains the date of birth, sex and marital status of all members of the household. Since age and sex are essential variables for the running of the survey (and also for certain processing stages such as weighting), they must be collected at all costs. In the rare cases where respondents refuse to give their age, the interviewer must estimate it.

The third block of the questionnaire, the *general component (H05)*, consists of questions related to disability during the past two weeks, use of health care services, activity limitations, chronic health problems and sociodemographic information (such as ethnicity, language, education level, employment status, income). All household members must complete this component.

The last block, the *health component (H06)*, includes questions dealing with access to health services, health status (vision, hearing, etc.), physical activities, use of tobacco, alcohol, medications, etc. These questions are asked to the selected person only. For more information on the questionnaire, see Statistics Canada (2003b).

Figure 5.1 Blocks of the questionnaire



5.2 Types of Content

The NPHS questionnaire also contains different types of content. *Core content* is defined as content that will be part of the questionnaire for all cycles of the survey. The subjects dealt with include disability in the past two weeks, use of health care services, activity limitations, chronic conditions, women's health, blood pressure, tobacco and alcohol. The core content appears in both the general component (H05) and the health component (H06) of the questionnaire. *Focus content* consists of questions that can be added to the survey for one cycle only or be repeated for several cycles of the NPHS. This content is requested from the selected person only (which means that it appears in the health component (H06)). Examples of focus content are stress in Cycle 1, access to health care services in Cycle 2 and personal health in Cycle 3.

Buy-in content is made possible where agencies external to Statistics Canada pay to add content to the regular survey. In the first three cycles of the NPHS, a few provinces took advantage of this opportunity to add content or more detailed questions. In Cycle 1, Manitoba and Alberta added content (on coping) specifically for their buy-in sample. In Cycle 2, the Health Promotion Survey (which had been conducted as a separate survey on health in Canada in 1994) was integrated into the NPHS and is therefore considered as buy-in content. In Alberta, extra information on health was collected from 12-17-year-olds; questions on coping were administered to respondents aged 18 and over; questions on attitudes toward parents were asked to 12-17-year-olds; questions on sexual health were asked to 15-59-year-olds; and the subject of personal safety/violence was

raised in all interviews which were not conducted by proxy. In Cycle 3, diets and nutrition as well as questions on smoking from the Health Promotion Survey were buy-in content.

There is a final type of content which is referred to as *supplemental content*. This is actually a derivative of the buy-in content, since pre-screening or filter questions that are incorporated into the NPHS will lead to an independent survey. In Cycle 1, the Health Promotion Survey (sponsored by Health Canada) was conducted on all longitudinal respondents aged 12 and over whose interviews had not been conducted by proxy. In Cycle 2, a supplementary survey on asthma was attached to the NPHS (sponsored by Health Canada). The persons selected for this survey had to state, in the NPHS interview, that they had been diagnosed as asthmatic by a health care professional. In addition, only longitudinal respondents were eligible for the pre-screening question and hence for the supplementary survey on asthma. In Cycle 3, a supplement on food insecurity (sponsored by Human Resources Development Canada) used the NPHS as the filter mechanism for an independent survey. The NPHS asked three filter questions in the Cycle 3 cross-sectional survey (with no age restriction), and the answers to these questions determined whether or not the respondent was selected for the food insecurity supplement.

5.3 Questions on Data Linkage and Sharing

Since the NPHS shares the data collected with its partners, it is essential to ask respondents whether they agree to share their information. For this purpose, a *data sharing question* is used. It reads as follows:

> *"To avoid duplication, Statistics Canada intends to share the information from this survey with provincial ministries of health (all cycles), Health Canada (all cycles), Employment and Immigration Canada (Cycle 1) and Human Resources Development Canada (Cycle 3). These organizations have undertaken to keep this information confidential and use it only for statistical purposes. Do you agree to share the information provided?"*

It should be noted that in each cycle, the list of partners with whom the information was shared varied slightly (see parenthetical notes in box).

Moreover, for linkage purposes, provincial health insurance numbers are also collected from respondents. This information is distributed on request to provincial health ministries, enabling them to link survey data with provincial administrative data related to health, such as physicians' billings and hospital admissions. In cycles 2 and 3, the NPHS requested a provincial health insurance number if it had not been supplied previously or if there had been a change.

In cycles 1, 2 and 3, the question on *permission to link* was requested as follows:

> "*We are seeking your permission to link information collected during this interview with provincial health information. This would include information on past and continuing use of services such as visits to hospitals, clinics, doctor's offices or other services provided by the province. This information will be used for statistical purposes only. Do we have your permission?"*

5.4 Questionnaire Testing

Before collection, it is important to test both the questionnaire and the computer application and to train the interviewers. The NPHS has always recognized the importance of testing, and in each of its cycles, two tests are conducted. In general, the first test takes place in November, in English only, and is primarily intended to verify that the computer application is functioning properly. The second test, which is conducted in both languages, checks the functioning of the computer application following the changes dictated by the first test.

6. Data Collection

Because of the complexity of the NPHS, collection procedures must be clearly established and fixed in time. The collection periods from one cycle to the next must be very similar, the collection mode and application must remain substantially the same, questionnaires must be changed as little as possible, etc. Furthermore, to ensure that collection is done properly, interviewers must be adequately trained and must follow the instructions dictated by the survey.

6.1 Collection Period

Each NPHS collection cycle is spread over a period of just over twelve months, which is divided into four periods (P1, P2, P3 and P4). The first period usually begins in June, with the subsequent periods starting in August, November and February respectively. Ideally, these periods should be fixed from one cycle to the next, but for operational reasons, the duration of each has varied slightly over the cycles. Table 6.1 shows the schedule used in each cycle. In each period, a pre-assigned part of the sample is contacted for an interview. Insofar as possible, longitudinal members are assigned to the same period from one cycle to the next.

Table 6.1: Schedule showing collection periods (P1, P2, P3, P4 and P5) for the first three cycles

Cycle	**Month of collection**													
1	***1994***							***1995***						
	June	***July***	***Aug.***	***Sept.***	***Oct.***	***Nov.***	***Dec.***	***Jan.***	***Feb.***	***Mar.***	***April***	***May***	***June***	***July***
	P1		P2			P3			P4				P5	
2	***1996***							***1997***						
	June	***July***	***Aug.***	***Sept.***	***Oct.***	***Nov.***	***Dec.***	***Jan.***	***Feb.***	***Mar.***	***April***	***May***	***June***	***July***
	P1		P2			P3			P4				P5	
3	***1998***							***1999***						
	June	***July***	***Aug.***	***Sept.***	***Oct.***	***Nov.***	***Dec.***	***Jan.***	***Feb.***	***Mar.***	***April***	***May***	***June***	***July***
	P1		P2		P3				P4			P5		

To convert cases of total non-response, the NPHS has a re-send procedure. Non-response cases that are obtained during the first period and unresolved are re-sent to the field in the third collection period. The same is true for cases from the second period which are re-sent in the fourth period. Lastly, a fifth collection period (P5) that usually takes place in June is used primarily to convert any remaining non-response cases accumulated during the four regular collection periods.

The fifth collection period is also the time for returning to the field any cases in which the respondent is listed as deceased but where the death could not be confirmed by vital statistics data. The NPHS does not consider a person as being deceased so long as the death has not been confirmed through matching with mortality files. Furthermore, the most recent version of the mortality file (which is two years out of date) is not available before March; matching with this file therefore cannot be done before April. Given the NPHS collection schedule, this time coincides with the start of the fourth collection period. Consequently, when the matching is performed, cases that did not match are re-sent to the field in the fifth collection period for interview purposes.

Finally, another type of case is sent to the field, namely respondents who have moved into a health care institution. If a longitudinal respondent moves into such an institution, the institutional component will be used as the data collection tool and will provide the necessary information to the household component of the NPHS. During the collection for the institutional component, it is possible that a longitudinal respondent will be observed to have gone back to living in a private dwelling. In this event, the case goes to the household component and is sent to the field during the fifth collection period. It should be noted that starting in Cycle 3, a respondent originally selected by the institutional component who moves to a private dwelling is neither covered nor interviewed by the household component of the NPHS.

6.2 Collection Mode

As noted in Section 4, the NPHS sample design mainly uses the LFS sampling frame. Therefore, the collection method for the first cycle of the NPHS is that chosen by the LFS, namely an initial contact in person. If the interview cannot be completed in person, the interviewer may offer the respondent the opportunity to complete the interview by telephone. At the end of the Cycle 1 interview, all respondents are asked their telephone number. They are also asked whether they prefer to be contacted by telephone in the next cycle. Notably, approximately 95% of respondents agree to have their next interview done by telephone. Table 6.2 shows the distribution of telephone interviews versus face-to-face interviews by component (general or health) for the panel and the cross-sectional sample.

Table 6.2: Percentage distribution of telephone interviews vs. face-to-face interviews by component, for each cycle

Component	Collection mode	Cycle 1		Cycle 2		Cycle 3	
		Panel	Cross-sectional	Panel	Cross-sectional	Panel	Cross-sectional
General component (H05)	*Telephone*	18	21	95	99	95	92
	Face-to-face	82	79	5	1	5	8
Health component (H06)	*Telephone*	25	28	95	99	95	91
	Face-to-face	75	72	5	1	5	9

6.3 Collection Aspects

The NPHS uses LFS interviewers to conduct interviews and thus has the benefit of experienced interviewers who need training only on the concepts of the survey and the questionnaire. In all cycles, interviewers receive training to inform them of changes made to the questionnaire since the last cycle. New interviewers (those who have never worked on the NPHS) receive more in-depth training.

One of the key features of a longitudinal survey is that contact is maintained with the members of the panel. If they move to a new dwelling or another province, they are still eligible to be interviewed. Thus, when longitudinal members move, every effort is made to locate them again. Indeed, tracing is an increasingly important component of interviewers' work.

To facilitate tracing, respondents are asked for information to help contact them in the next cycle. They provide the name, address and telephone number of two persons whom the NPHS could contact to trace the longitudinal respondent if he/she moves. This information enables the NPHS to continue collecting data from longitudinal respondents. In each cycle, approximately 2% of longitudinal respondents move. It is therefore very important, throughout the survey, to be able to contact the panel members.

Respondents who move into a health care institution are interviewed by the NPHS institutional component. A number of survey questions are the same for the institutional and household components, making it possible to collect longitudinal information for these respondents.

Efforts are also made to minimize non-response. Interviewers are specially trained, and specialized or experienced interviewers handle difficult cases. If a household refuses to respond, a letter is sent to explain the authenticity of the survey and the importance of responding to it.

The NPHS uses computer-assisted interviewing for its first three cycles. The CASES programming language is used in the collection application. An initial edit of the data collected is done online, during the interview. The edit process serves to bring out obvious errors in the data, identify inconsistencies and make various corrections, confirming the answer directly with the respondent.

To conduct the edit, edit rules associated with computer-assisted interviewing are programmed in advance into the application. The NPHS distinguishes three types of edit rules:

i) Validity rules: These rules verify that only acceptable and valid answers are given to each question. They also ensure that the values provided lie within a plausible range.

ii) Consistency rules: These rules verify that known or expected relationships between several questions (variables) are not contradictory. For these rules, two or more variables are checked for the same respondent or several respondents.

iii) Longitudinal rules: These rules ensure that relationships between questions (variables) are not contradictory from one cycle to the next for longitudinal panel respondents. These rules compare the same variable for all cycles to make sure that the data are consistent. For this edit, historical data are downloaded into the collection application in each cycle.

Where there is an invalid or inconsistent response, the system posts a prompt and the interviewer confirms the difference with the respondent. There are two categories of online edit rules: hard and soft. Hard edit rules ensure that pre-determined conditions are met. For example, a hard edit rule checks whether the value entered in response to a question lies between predetermined minimum and maximum values or corresponds to a valid category. If the value entered in response to a question does not satisfy the hard edit rule, a prompt appears on the screen and the interviewer must enter a valid response to be able to continue the interview.

In turn, a soft rule is used where a possible but unusual response is entered. If the response does not satisfy the soft rule, a prompt appears on the screen and the interviewer must confirm the unusual answer with the respondent. The interviewer must then change or accept the response to be able to continue the interview.

6.4 Proxy interviews

The general component is completed for all members of the household. The information requested may be provided by any adult household member who knows the answers. The health component must be completed by the selected person only. However, proxy interviews are allowed in four situations: i) the interviewer can confirm that the respondent will not be present during the entire collection period, ii) respondent has a mental or physical incapacity preventing the interview, iii) there is a language barrier, and iv) the respondent is under 12 years of age. It should be noted that in Cycle 1, respondents under 12 years of age were interviewed by the NLSCY. Even when responding by proxy is authorized, some parts of the health component cannot be completed by a person other than the selected respondent. These parts, including questions on physical activities and mental health, are passed over in a proxy interview. Table 6 shows the percentage of interviews conducted by proxy in the different cycles of the survey.

Table 6.3: Proxy interviews rates by component and cycle

Component	Age	Cycle 1		Cycle 2		Cycle 3	
		Panel	Cross-sectional	Panel	Cross-sectional	Panel	Cross-sectional
General component (H05)	All	n.s.*	54.7	26.1	54.8	22.0	54.0
	12 and over	30.3	43.6	18.0	45.1	14.9	42.7
	Under 12	n.s.*	n.r.*	100.0	100.0	100.0	100.0
Health component (H06)	All	n.s.*	n.a.*	11.8	12.4	10.9	13.7
	12 and over	4.2	4.2	2.0	2.3	2.7	2.4
	Under 12	ns.*	n.a.*	100.0	100.0	100.0	100.0

* In Cycle 1, respondents under 12 years of age were interviewed by the NLSCY.
n.a.: not applicable.
n.s.: not stated.

7. Data Processing

Figure 7.1 shows the different stages whereby the data collected are processed to create the final files (these will be discussed in Section 9). In chronological order, these include unpacking, verify, reformat, coding, edit, production of derived variables, weighting, and creation of master files. These main processing stages also interact with the address register and the data dictionary. All members of the NPHS project team are actively involved in one or more of these processing stages.

Subsection 7.1 provides an overview of the process surrounding data collection and transmission to Head Office, where the processing stages are carried out. The following subsections successively detail each of the processing stages illustrated in Figure 7.1.

7.1 Data Collection and Transmission

To collect the survey data, each interviewer is equipped with a portable computer in which he/she directly enters the interviewee's answers. As discussed in section 6.3, some edit rules are automated in the computer application, thereby improving data quality. Once collected, the data are transmitted to the interviewer's regional office, then retransmitted to Head Office in Ottawa, where the final data are processed.

7.2 Unpacking

The data are transmitted in the form of encrypted packets to ensure respondent confidentiality. The first stage of processing is the unpacking of the data. This involves unpacking the data packets to create files in dbf format and then copying them to several files in text format (txt). In this stage, files are organized into a readable format (dbf or txt). For illustration purposes, Figure 7.2 shows the process of unpacking packets of compressed and encrypted files into dbf or txt files.

7.3 Verify

At this stage of processing, a number of validations are done at the record level. The first step is to make sure that the files contain no duplication. The next step is to validate the age and sex variables by identifying any missing or invalid values. The age and sex variables in the current cycle are compared to those in previous cycles. Where differences are observed, the notes are analysed to determine whether a change is needed. It is also at this stage that the final response code is derived by using the response code received from collection and performing automated and manual edits. Response codes will play a crucial role in subsequent stages (in weighting, for example).

Figure 7.1 Processing of NPHS data

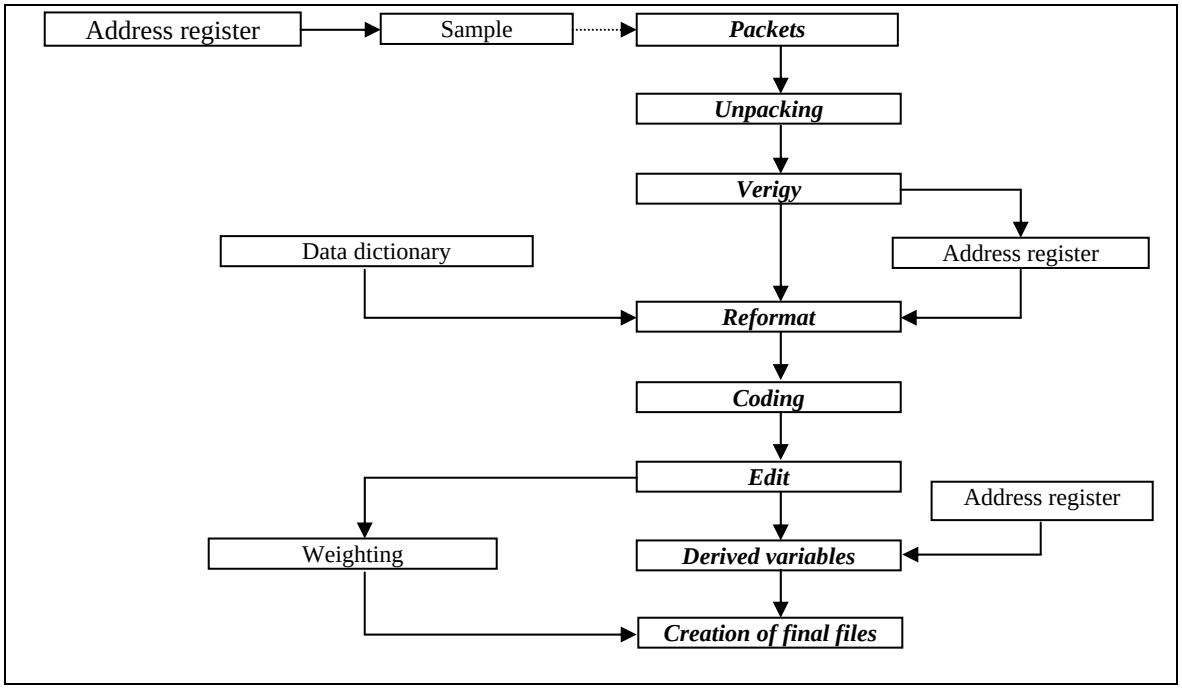
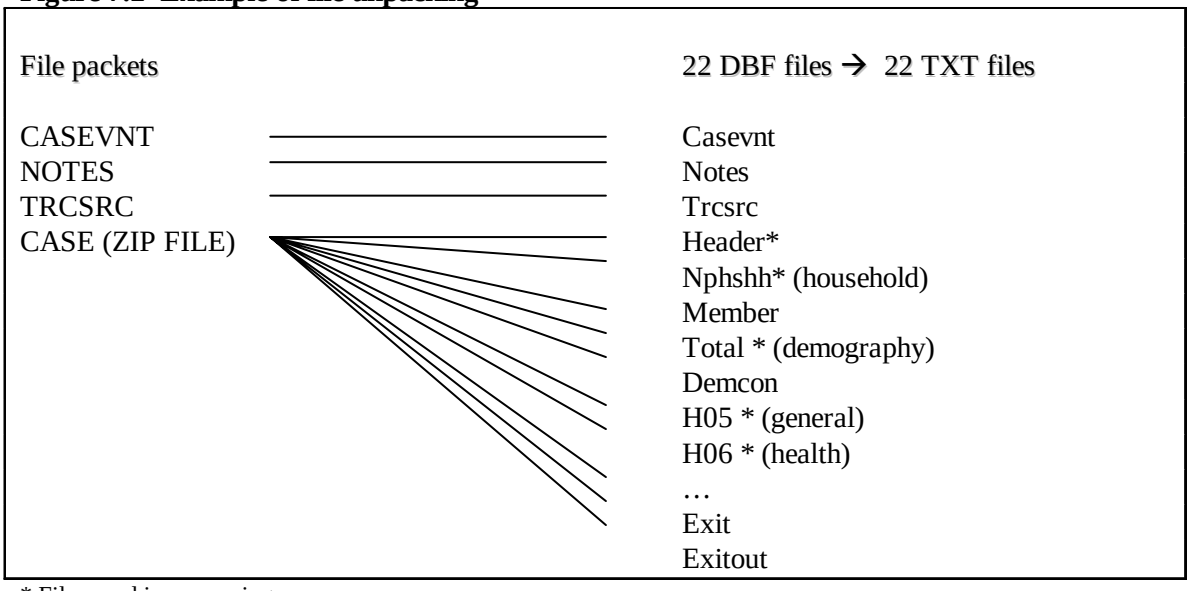


Figure 7.2 Example of file unpacking



* Files used in processing.

7.4 Reformat

Data reformatting is an iterative process applied to variables until the desired result is obtained. Such an iterative process takes place between the NPHS processing team that edits the results obtained and the team of programmers who make the necessary adjustments to programs until the expected results are obtained. One of the main activities in this process is to take the codes received from the application and recode them according to standard non-response categories. The standard codes used by the NPHS vary depending on the length of the variable's field; they are: not applicable (6, 96, 996, …), don't know (7, 97,997, …), refusal (8, 98, 998, …), and not stated (9, 99, 999, …). In addition, all categories of all answers to questions with the instruction "mark all that apply" are converted to dichotomous "yes or no" variables.

During the data reformatting process, the data are also prepared for the next stage, namely coding. But reformatting also includes creating files made up of long answers. After this first recoding stage, some variables are coded to "not applicable" to ensure that plausible values are properly processed according to the skip rules for the questionnaire. In this process, new household (H3HH), member (H3MM), general (H05) and health (H06) files are created.

7.5 Coding

In a survey, coding is the operation by which codes are assigned to long answers. This stage makes it easier, down the line, to analyse the survey results. In the NPHS, this process is used for various questions for which a long answer could be given. The variables on activity limitations, medications and employment status were coded using international classification systems. Figure 7.3 shows the classification systems used by the NPHS in cycles 1, 2 and 3.

Figure 7.3 Classification systems used by the NPHS

Variable		Classification system
Activity limitation	————	International Classification of Diseases, Version 9 (ICD-9)
Employment status	————	North American Industry Classification System (NAICS)
Medications	————	ATC Classification System for Human Medicines, Canadian Edition

Lastly, for all other variables accepting answers other than those predetermined in the application, these answers are submitted to coding to form new separate categories or classify them to an existing category.

7.6 Edit

During the interview with the respondent, an online edit is performed to identify obvious errors, identify inconsistencies and make various corrections online (see subsection 6.3). Unfortunately, not all the edit rules can be incorporated into the collection instrument, and some must be applied at Head Office at this point in the processing (head office edit rules). Unlike the online edit rules, no correction is made if the head office edit rule is not satisfied. Inconsistencies are left

unchanged but are recorded. There is only one exception to this procedure: rules concerning relationships between respondents within the same household. If an inconsistency is observed in these relationships, a manual correction is made to correct errors. However, the edit rules on relationships are applied separately from other rules.

As in the case of online edit rules, there are three types of head office edit rules: validity rules, consistency rules and longitudinal rules.

In general, edit rules are applied independently for each section of the questionnaire: household (H3HH), member (H3MM), general component (H05) and health component (H06). However, it sometimes happens that rules will involve more than one section of the questionnaire. Furthermore, as noted above, longitudinal rules use more than one data cycle (for example, validation of the provincial health insurance number involves the use of data from previous cycles).

7.7 Derived Variables

The NPHS provides its users with a whole range of derived variables. These variables, created from items in the NPHS questionnaire, are attractive for several reasons: i) they facilitate data analysis; ii) they serve to protect respondents' confidentiality; and iii) they promote consistency between different analyses. In some cases, a derived variable is obtained merely by grouping response categories of an existing variable. In other cases, several variables are combined to form a new variable.

Derived variables may be defined differently from one cycle to the next, but in general, an effort is made to ensure stability. If the definition of a derived variable is changed, this variable is again derived for the previous cycles in the longitudinal file. The NPHS has derived variables concerning geography (e.g., urban/rural area), income (derived income adequacy – 2 groups), household (household size, marital status – grouped), limitation of activity (a flag), chronic health problems (number of problems), education level (highest level of education), physical activity (physical activity index) and various other subjects. Information concerning the derivation of these variables is provided with the microdata files.

7.8 Weighting

The principle behind estimation in a probability sample such as in the NPHS is that each person in the sample "represents," besides him/herself, several other persons in the population. In the weighting phase, an associated weight is determined for each sampling unit. In the NPHS, this weighting is complex. The NPHS uses multiple survey frames (see Section 3); it provides both longitudinal and cross-sectional estimates; and the survey includes buy-in and top-up samples. The weighting method used by the NPHS is described in detail in Section 10.

7.9 Creation of Final Files

The result of all the processing steps, is in fact, the creation of different files. This steps consists of combining the processed data, the derived variables and the sampling weights. For further details on the list of products, see Section 9.

8. Non-response

The NPHS, like any voluntary survey, must deal with non-response. Usually two types of non-response are identified: *total non-response* and *item non-response* (also commonly known as *partial non-response*). The effect of non-response on the survey results is a source of non-sampling error in surveys; it therefore merits attention. Subsection 8.1 examines total non-response and section 8.2 looks at item non-response.

8.1 Total Non-response

Since the NPHS was both a cross-sectional and a longitudinal survey during its first three cycles, and since the concept of non-response is not defined in exactly the same way in both cases, we will first look at total non-response for cross-sectional NPHS samples in subsection 8.1.1 and then total non-response from a longitudinal standpoint in subsection 8.1.2.

8.1.1 Cross-sectional Samples

In defining non-response in cross-sectional samples, it is important to distinguish between the two components: the general component and the health component (see subsection 5.1 for further details). For the general component, each member of the household interviewed is assigned a response status: complete response, partial response or non-response. The status assigned depends on the level of completion of the questionnaire. Critical points in the questionnaire are established in advance to determine partial or complete responses. A person will be considered non-respondent to the general component if he/she has not answered any question in that component. The person will have partial response status if at least one question has been answered but the person has not gone beyond a predetermined question located approximately midway through the general component. The person will have complete response status if going beyond that question. Table 8.1 shows the questions that are considered critical points in the questionnaire and which serve to distinguish partial responses from complete responses. Since the makeup of the questionnaire may vary from one cycle to another, the critical points are presented independently for each of the cycles.

The same principle applies to the health component, where each selected person is assigned a response status according to how far he/she advanced through the questionnaire. Table 8.1 shows the questions used in each cycle to determine partial or complete responses to the health component. Cycle 2 is a special case, owing to the fact that in that cycle, Alberta bought additional content to be administered exclusively to children. Thus, for the children selected in that province, the critical point in the questionnaire was adjusted to more fairly represent the concept of partial or complete response.

It should be noted that for the general component, the data file disseminated to users and partners contains only the records of persons with complete response status in the general component. The same is true for the health component file. Also, since a response to the health component is conditional on having first responded to the general component, all respondents in the health component file may be considered to have supplied complete responses to the general component. This is an important point, since the data file for the health component contains not only the variables collected in the health component but also data obtained from selected persons in their interviews for the general component.

In summary, for the cross-sectional aspect of the NPHS, persons providing partial responses are considered part of total non-response along with people who did not respond to the questionnaire. There are many reasons explaining these non-responses, such as a prolonged absence from the household or a simple refusal to participate in the survey.

To compensate for this loss of sample due to total non-response, and to control the potential bias that may exist owing to the fact that the profile of respondents may be different from that of non-respondents, the sampling weight of respondents is adjusted. Section 10 deals with weighting and the techniques used to make this adjustment.

Table 8.1: Questions used to determine response status in the general and health components

Cycle	Component	Questions determining response status
1	General	*In what country were you born?* – SOCIO-Q1
	Health	*In the past 12 months, did you have any injuries serious enough to limit your normal activities?* – INJ-Q1
2	General	*In which languages can (name) conduct a conversation?* – SOCIO-Q5
	Health	*Are you usually able to see well enough to read ordinary newsprint without glasses or contact lenses?* – HS-Q1
	Health – Children in Alberta	*Does (name) have asthma that has been diagnosed by a health professional?* – KCHR-Q4
3	General	*To which ethnic or cultural group(s) did (name)'s ancestors belong?* – SOCIO-Q4
	Health	*Are you usually able to see well enough to read ordinary newsprint without glasses or contact lenses?* – HS-Q1

8.1.2 Longitudinal Sample

As described in Section 4, the longitudinal sample consists of 17,276 selected persons who at least had a complete response status in the general component of Cycle 1.

In each cycle, each member of the panel is assigned a response status relating to that cycle. Depending on the outcome of the interview, a person is assigned one of the following five statuses: non-respondent, partial respondent, complete respondent, deceased, in institution. Similarly to what is done for cross-sectional samples, there is a series of rules for distinguishing between these statuses. If the longitudinal member does not have at least complete response status in the general component, that person is considered a non-respondent for longitudinal purposes. A person who completed the general component but did not respond or only partially responded to the health component is considered a partial respondent. A person who provided complete responses to both components is considered a complete respondent. Lastly, the status "in institution" indicates that a complete response was obtained via the collection mechanism for the institutional component.

Over the cycles, the response statuses of a given respondent are concatenated into a single variable called the "longitudinal response pattern." This variable is available in the longitudinal files and can be used to rapidly obtain the response profile of a member of the panel. This variable is also used to identify different analytical subsets as described in Section 9.

In a longitudinal survey, non-response to one or more cycles is quite costly. In a sense, it breaks the chronological sequence of information available about a respondent, which may make it more

complex to conduct analyses on the data. Worse yet, chronic non-response to the survey reduces the sample size available for analysis, thus diminishing the potential for observing statistically significant results. Consequently, many efforts have been made in the NPHS to minimize non-response (see subsection 6.3 for further details).

8.1.3 Response Rates

Table 8.2 shows the response rates for the cross-sectional sample and the panel, by cycle. For the cross-sectional sample, the rate reported for the general component is actually a response rate calculated at the household scale. Thus, the rate represents the percentage of households that agreed to participate in the survey out of all those contacted. In the vast majority of cases, when a household agrees to participate, the information for the general component is obtained for all members of the household. The rate for the health component represents the percentage of selected persons who responded completely to this component.

For the panel, the rates reported for cycles 2 and 3 represent the proportion of the 17,276 respondents who responded completely in the interview. The Cycle 1 rate represents the proportion of complete responses to the health component among the persons originally selected for this cycle. Note that deceased persons or persons in institutions are considered respondents.

For more details on response rates, please see the user guides provided with the data for each cycle (Statistics Canada, 1995, 1998a, 2000 and 2002).

Table 8.2: Response rates by cycle according to panel and cross-sectional sample

Cycle	Panel	Cross-sectional sample	
		General (households)	Health (persons)
Cycle 1	83.6	88.7	96.1
Cycle 2	92.8	82.6	95.6
Cycle 3	88.2	87.6	98.5

8.2 Item Non-response

When information is available for some questions only, such as when a person answers only part of the questionnaire, this is known as item non-response. When discussing the content of data files, these non-responses are often referred to as missing values. In the NPHS, there are four categories of missing values:

- Not applicable: Indicates that the question did not apply to the respondent.
- Don't know: Indicates that the respondent was unable to answer the question.
- Refusal: Indicates that the respondent refused to answer the question.
- Not stated: Indicates that the information could not be reported because the question was not asked but should have been asked. Typically, this occurs when a question is associated with a filter and the filter question was not answered, thus preventing the interviewer from obtaining answers to the questions conditional on this filter. It should be noted that all variables for deceased persons are marked "not stated" starting in the cycle in which they are identified as being deceased.

Note that missing values are recoded in data files, using a coding system widely used in household surveys at Statistics Canada (see subsection 7.4).

A popular approach in the survey field is to impute missing values, that is, to assign a replacement value for a missing value. In the case of the NPHS, there is no imputation, and therefore it is up to analysts to decide how to deal with missing values.

9. Different Products

Several types of data files are produced using the data collected by the NPHS. The main differences between these files are the cross-sectional or longitudinal aspect and the number of variables and observations included.

Up to and including Cycle 3, two cross-sectional data files are produced in each cycle, one for each component (general and health). The file for the health component also contains the variables from the general component, for the respondent selected from the household. Other files are also produced when special themes or supplements are added to the main survey. In Cycle 1, there was a supplement to the NPHS (a special survey) called the Health Promotion Survey (HPS). The HPS questionnaire was administered to persons aged 12 and over in the panel. In Cycle 2, a supplement on asthma was added for both the general component and the health component. In Cycle 3, a supplement on food insecurity was added to the health component.

Starting with NPHS Cycle 2, it became possible to produce longitudinal files. There are three types of longitudinal files: *square, full* and *partial*. The *square file* contains data for all 17,276 members of the longitudinal panel, regardless of the status of their questionnaire (complete, partial, non-response, etc.). The number of records included in this file is therefore identical from one cycle to the next, but the number of variables increases after each cycle of the survey, since after each cycle, the data collected are added to those from previous cycles. The *full file*, in turn, contains only respondents for whom there are complete responses for the current cycle and all previous cycles. Therefore, the number of records contained in this file can only decrease over time. It should be kept in mind that persons who are deceased or living in institutions are considered to be respondents. A *partial file* was created in Cycle 2 and contained all persons having either a complete or partial response to the first two cycles of the survey.

Since the data are used by a variety of users, different versions of these files must be produced to allow data sharing with other organizations or survey partners and to make sure that data confidentiality is maintained. Three versions of files exist in the NPHS: *master files*, *share files* and *public use microdata files* (*PUMFs)*.

Master files contain all respondents with all the variables in the survey. Only variables serving to identify the respondent directly (name, full address, etc.) have been removed. These files are used by deemed employees of Statistics Canada.

Share files contain the same variables as the master files but exclude respondents who refused to have their data shared with other organizations that are subject to the Statistics Act (i.e., organizations that guarantee data confidentiality). To obtain the list of organizations or partners with which the information collected is shared, see subsection 5.3. It is worth noting that usually around 95% of respondents agree to share their data.

Lastly, PUMFs, like master files, contain all respondents but exclude variables that might serve to identify respondents either directly or indirectly. These files are accessible, at a cost, to everyone requesting them. Section 12 provides more information on how these PUMFs are created.

There is another type of file in addition to those described in the preceding paragraph: the *linkage file*. This file contains the subset of respondents who agreed to share their data and also those who agreed to their data being linked (usually this amounts to approximately 90% of respondents). The file contains direct identifiers that can be used in a linkage process, as well as the provincial

health insurance number supplied by the respondent when agreeing to the linkage of his/her data. The linkage file is available only to provincial health ministries (on request).

Cross-tabulating all these characteristics generates a sizable number of files, produced in each cycle. Also, each cycle has its own distinctive features and presents a number of exceptions. Table 9.1 summarizes the various data files produced for each of the cycles. Bootstrap weights files and dummy files are also shown in the table; they will be described in subsections 11.1.1.2 and 12.3 respectively.

Table 9.1: The various files produced as part of the NPHS

File	Cycle 1		Cycle 2		Cycle 3	
	Weighted data file	Bootstrap weights file	Weighted data file	Bootstrap weights file	Weighted data file	Bootstrap weights file
	Number of obs.	Number of weights	Number of obs.	Number of weights	Number of obs.	Number of weights
Master files						
Cross-sectional:						
H35 – general	58,439	500	210,377	500	49,046	500
H356 – health	17,626	500	81,804	500	17,244	500
H357 – HPS	13,400	1,000	unavail.	unavail.	unavail.	unavail.
H35 – asthma	n.a.	n.a.	2,364	2,000	unavail.	unavail.
H356 – asthma	n.a.	n.a.	1,102	2,000	Unavail.	unavail.
H356 – food insecurity	n.a.	n.a.	unavail.	unavail.	Unavail.	unavail.
Longitudinal:						
LNGF – full	n.a.	n.a.	15,670	500	14,619	500
LONG –square	unavail.	unavail.	17,276	500	17,276	500
LNGP – partial	n.a.	n.a.	16,168	500	unavail.	unavail.
Share files						
Cross-sectional:						
H35 – general	55,613	500	196,658	500	47,632	500
H356 – health	17,011	500	77,403	500	16,787	500
H35 - asthma	n.a.	n.a.	1,814	2 000	n.a.	n.a.
H356 - asthma	n.a.	n.a.	828	2 000	n.a.	n.a.
H356 – food insecurity	n.a.	n.a.	n.a.	n.a.	1,265	1,000
Longitudinal:						
Complete	unavail.	unavail.	14,860	500	14,250	500
ID files						
Cross-sectional:						
H35 - general (share)	55,613	500	unavail.	unavail.	unavail.	unavail.
H356 – health (share)	17,011	500	70,980	500	16,450	500
Longitudinal:						
LNGF – full (master)	n.a.	n.a.	12,693	500	11,611	500
LNGF – full (share)	n.a.	n.a.	12,495	500	11,487	500
Dummy files						
Cross-sectional:						
H35 - general	58,439	500	210,377	500	49,046	500
H356 – health	17,626	500	81,804	500	17,244	500
Longitudinal:						
LNGF – full	n.a.	n.a.	15,670	500	14,619	500
PUMFs						
Cross-sectional:						
H35 - general	58,439	unavail.	210,377	unavail.	49,046	unavail.
H356 – health	17,626	unavail.	81,804	unavail.	17,244	unavail.
H357 – HPS	13,400	1,000	unavail.	unavail.	unavail.	unavail.

Note:
- unavail.: not available
- n.a.: not applicable
- H35 - general: Cross-sectional file containing data from demographic (H03) and general (H05) components for all household members.
- H356 - health: Cross-sectional file containing data from demographic (H03), general (H05) and health (H06) components for selected persons.
- H357 - HPS: Cross-sectional file containing data from demographic (H03), general (H05) and Health Promotion Survey components for selected persons.
- H35 - asthma: Cross-sectional file containing data from demographic (H03) and general (H05) components for all household members in the asthma supplement.
- H356 - asthma: Cross-sectional file containing data from demographic (H03), general (H05) and health (H06) components for selected persons in the asthma supplement.
- H356 – food insecurity: Cross-sectional file containing data from demographic (H03), general (H05) and health (H06) components for selected persons in the supplement on food insecurity.

10. Weighting and Estimation

This section provides details on the weighting of the first three cycles of the NPHS. Subsection 10.1 deals with cross-sectional weighting, while subsection 10.2 concerns longitudinal weighting. Figures 10.1 to 10.6 schematically represent all the steps required to create each set of weights developed for the NPHS. For further details, see the public use microdata file user's guide (Statistics Canada, 1998a) or Stukel, Mohl and Tambay (1997) for Cycle 2. Table 10.1 lists the sets of weights for the NPHS.

Table 10.1 – Sets of weights for NPHS

	Cycle 1		Cycle 2		Cycle 3	
	Master	**Share**	**Master**	**Share**	**Master**	**Share**
Cross-sectional						
General	WT5	SHRWT5	WT56	WT56_S	WT58	WT58_S
Health	WT6	SHRWT6	WT66	WT66_S	WT68	WT68_S
Health (children + HPS)	n.a.	n.a.	WT66_N	WT66_SN	n.a.	n.a.
Longitudinal						
Square	WT64LS	n.a.	WT64LS	n.a.	WT64LS	n.a.
Full	n.a.	n.a.	WT66LF	WT66SLF	WT68LF	WT68_SLF
Partial	n.a.	n.a.	WT66LP	n.a.	n.a.	n.a.

n.a.: not applicable

10.1 Cross-sectional Weighting

Figures 10.1 to 10.5 schematically represent all the steps required for the cross-sectional weighting of the NPHS. Each adjustment shown in the figures is followed by a number in parentheses. This number is used in subsection 10.1.1 in describing in detail the different adjustments. A number in bold means that some aspects of the adjustment are specific to the cycle concerned.

Cycle 1 cross-sectional weighting (Figure 10.1):

Figure 10.1 summarizes the cross-sectional weighting for Cycle 1. The starting point is the LFS basic weight, or the ESS basic weight for Quebec. Different adjustments are made successively to obtain two sets of Cycle 1 weights: weight WT5 for the general component and weight WT6 for the health component.

In the first cycle, a few provinces purchased additional sampling units. In most cases, the weighting method was not affected since the additional units were selected in the same way as the units in the core sample. However, in British Columbia, the majority of the additional units were selected via the random digit dialling (RDD) method. These additional units obtained by RDD were located in the subprovincial region of Prince George. Three strata were thus sampled simultaneously in two different ways. The selection method for the RDD supplemental sample is different, and therefore, so is the weighting.

Cycle 2 cross-sectional weighting (figures 10.2 and 10.3):

The Cycle 2 weighting method is quite similar to that used in the first cycle. Firstly, "stripped" weights are obtained. The procedure followed for these weights is the same as for the first cycle (taking the Cycle 1 sample design (LFS and ESS) into account). However, some specific corrections in Cycle 1 are changed or eliminated, owing to the fact that the supplemental sample in Cycle 1 was not followed in Cycle 2. These weights are then adjusted for specific aspects of Cycle 2 to obtain the final weights for that cycle. Figures 10.2 and 10.3 summarize the cross-sectional weighting in Cycle 2.

Once again, some provinces purchased additional sampling units in order to produce reliable estimates at the scale of subprovincial regions. This time, supplemental samples were selected using RDD in Ontario, Manitoba and Alberta. The inclusion of these supplemental samples requires a series of additional adjustments, notably the integration of the two survey frames used. This results in the creation of three sets of weights: WT56 for the general component and WT66 and WT66_N for the health component.

WT66 is the weight used for analysing most variables in the health component. For the variables obtained from the HPS and questions on children's health care services, weight WT66_N is used instead. Weight WT66 applies to all age groups and all provinces. However, it does not cover the 0 to 11 age group in the same way for each province. In Manitoba and Alberta, for households selected via RDD, a child aged 0 to 11 was selected to answer the health component (in addition to an adult aged 12 and over). Thus for these provinces, WT66 covers children aged 0 to 11. For the other provinces, the only children included are those selected in Cycle 1 (they are therefore between 2 and 11 years of age). Additional details concerning the difference between WT66 and WT66_N are available in point (20) of subsection 10.1.1.

Cycle 3 cross-sectional weighting (figures 10.4 and 10.5):

Just as in Cycle 2, the Cycle 3 weighting method is largely based on that of Cycle 1. The starting point is the same as in Cycle 1 and a few basic adjustments are made to obtain stripped weights. New adjustments specific to the third cycle are then made to obtain the Cycle 3 final weights. Figures 10.4 and 10.5 summarize the cross-sectional weighting in Cycle 3.

What distinguishes Cycle 3 from the previous two cycles is the selection of a top-up sample. The top-up sample contains all dwellings that were non-respondent in Cycle 1 as well as an LFS subsample that includes children born in 1995 or thereafter and immigrants admitted to Canada since 1995, which are two sub-populations absent from the longitudinal sample for the year 1998-1999. Also, there was no buy-in of samples by the provinces for the third cycle.

The top-up sample of dwellings that were non-respondent in the first cycle was simply integrated into the core sample, since there were "stripped" basic weights in the first cycle for the units in the top-up sample. The other part of the top-up sample (sample of households including children born in 1995 or thereafter or recent immigrants) consisted of households initially absent from the 1994 population. The new households could therefore be simply added to the existing sample without changing their weight.

Two sets of weights were produced for the cross-sectional part of Cycle 3: weight WT58 for the general component and weight WT68 for the health component.

Weights specific to share files:

To obtain weights specific to the share files, an additional adjustment must be made to the weights in the master files in order to redistribute the weight of respondents who refused to share their responses to those who agreed to do so. Because of the small number of respondents who refuse to share their responses, this adjustment consists merely of making a new poststratification, using the same province-age-sex groups as for the initial poststratification (see point *(10)*).

Figure 10.1: Cross-sectional weighting - Cycle 1

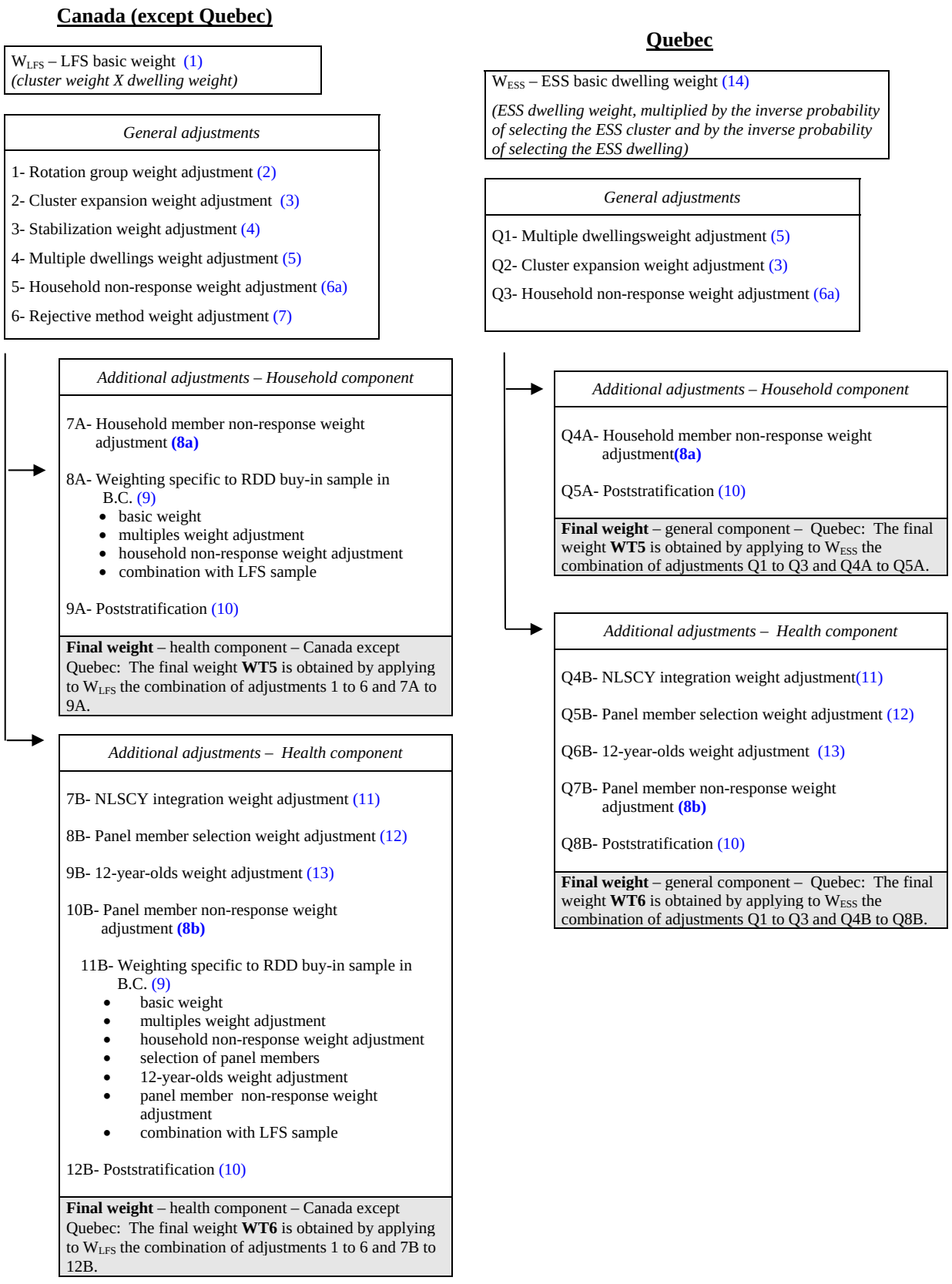


Figure 10.2: Cross-sectional weighting - Cycle 2 - Canada (except Quebec)

W_{LFS} – LFS basic weight (1)
(cluster weight X dwelling weight)

Obtaining stripped weights from Cycle 1 (adjustments retained from Cycle 1)

1- Rotation group weight adjustment (2)

2- Cluster expansion weight adjustment (3)

3- Stabilization weight adjustment (4)

4- Multiple dwellings weight adjustment (5)

5- Household non-response weight adjustment **(6a)**

6- Rejective method weight adjustment (7)

Adjustments specific to buy-in samples purchased by three provinces in Cycle 2

Basic weight (random digit dialling) (18)

B1- Manitoba sub-sampling weight adjustment (18a)

B2- Weight adjustment for telephone numbers selected twice (18b)

B3- Household non-response weight adjustment **(6b)**

B4- Weight adjustment for households with no telephone (18c)

B5- Multiple telephone lines weight adjustment (18d)

Additional adjustments – Household component

Adjustments specific to core sample:

7A- Household non-response weight adjustment **(6b)**

8A- Interprovincial migration weight adjustment(15)

9A- Initially absent cohabitants weight adjustment (16)

10A- Household member non-response weight adjustment **(8a)**

Adjustments specific tobuy-inl samples purchased by three provinces in Cycle 2:

B6A- Collective dwellings weight adjustment (18e)

B7A- Household member non-response weight adjustment **(8a)**

11A- Weight adjustment for integration of core sample and supplemental samples (19)

12A- Poststratification **(10)**

13A- Noise factor (17)

Final weight – general component – Canada except Quebec: The final weight **WT56** is obtained by applying to W_{LFS} the combination of adjustments 1 to 6 and 7A to 10A (core sample), adjustments B1 to B5 and B6A to B7A (buy-in samples), and finally by making adjustments 11A to 13A

Additional adjustments – Health component

Adjustments retained from Cycle 1(core sample only:

7B- NLSCY integration weight adjustment**(11)**

8B- Panel member selection weight adjustment (12)

9B- 12-year-olds weight adjustment (13)

New Cycle 2 adjustments (core sample only):

10B- Panel member non-response weight adjustment**(8b)**

11B- Interprovincial migration weight adjustment(15)

Adjustments specific to buy-in samples purchased by three provinces in Cycle 2:

B6B- Panel member selection weight adjustment (12)

B7B- Collective dwellings weight adjustment (18e)

B8B- Panel member non-response weight adjustment **(8b)**

12B- Weight adjustment for integration of core sample and buy-inl samples (19)

13B- Poststratification**(10)**

Final weight – health component – Canada except Quebec:

The final weight for the health sample is obtained by applying to W_{LFS} the combination of adjustments 1 to 6 and 7B to 11B (core sample) and adjustments B1 to B5 and B6B to B8B (buy-in samples), and finally by making adjustments 12B to 13B.

Two sets of weights were produced for the health component in Cycle 2 (see (20) for more details):

- **WT66**: to be used with the health component file for analysing most variables
- **WT66_N**: to be used for variables from the HPS or questions on children's health care services

Figure 10.3: Cross-sectional weighting - Cycle 2 - Quebec

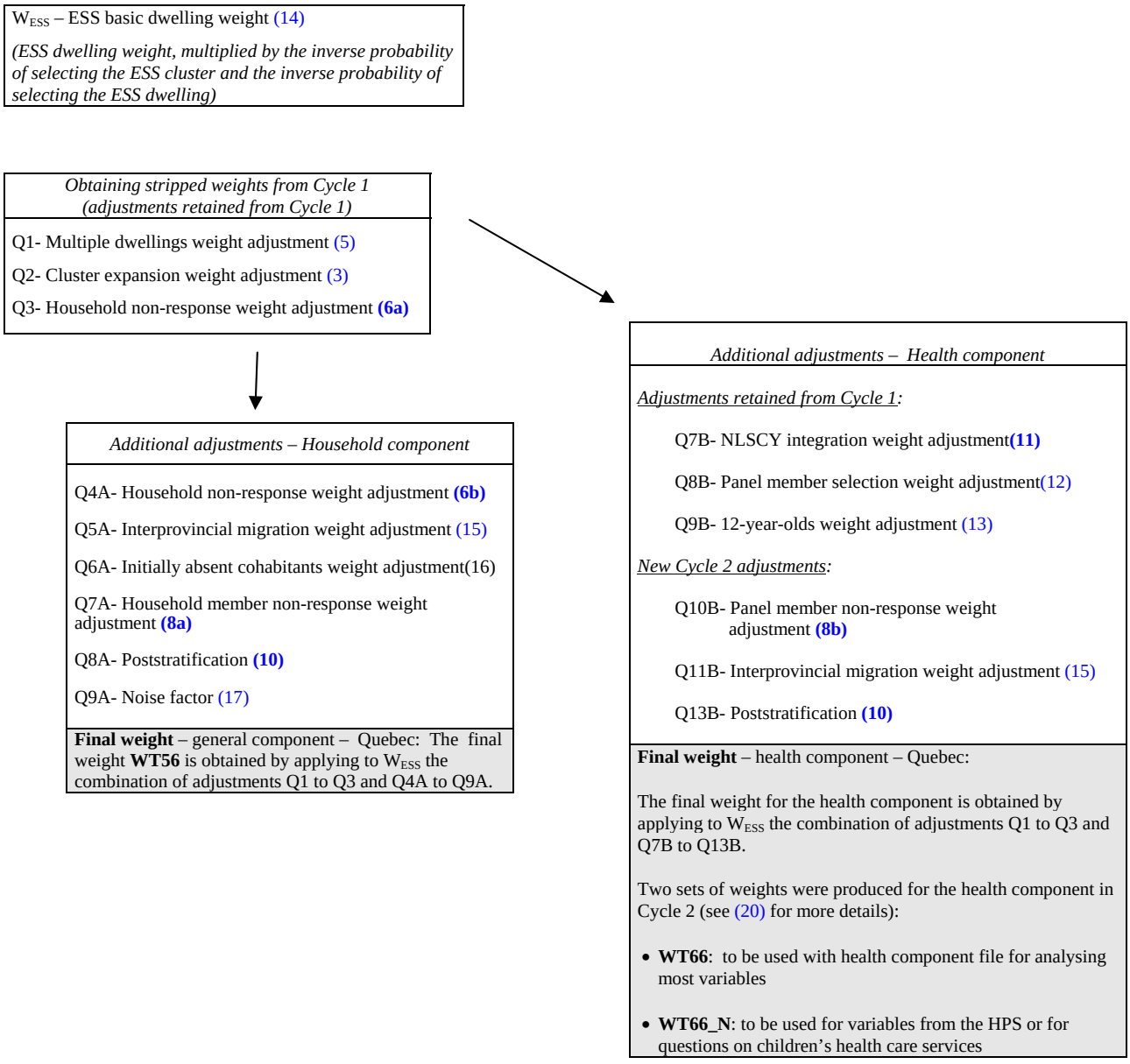


Figure 10.4: Cross-sectional weighting - Cycle 3 - Canada (except Quebec)

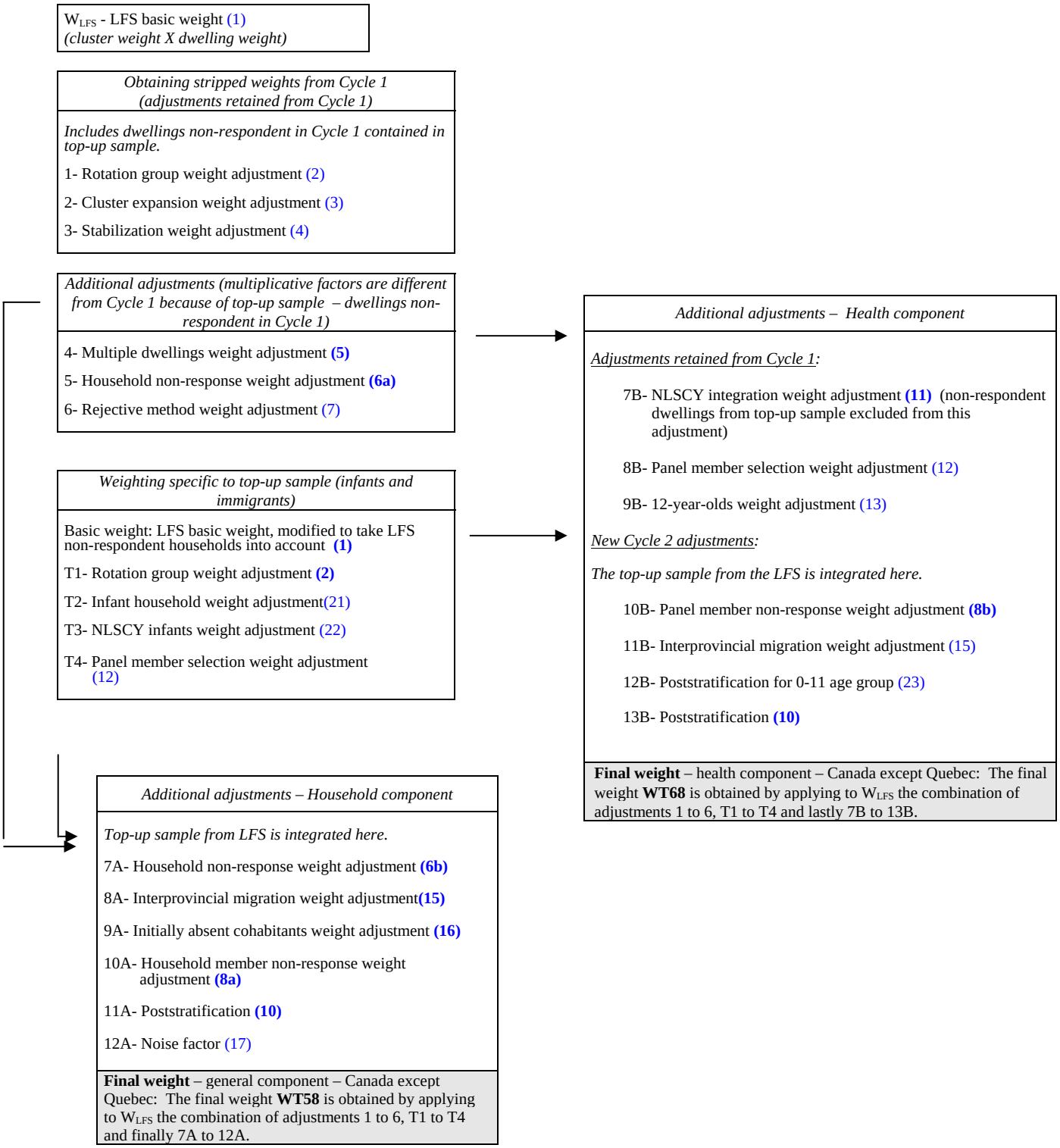
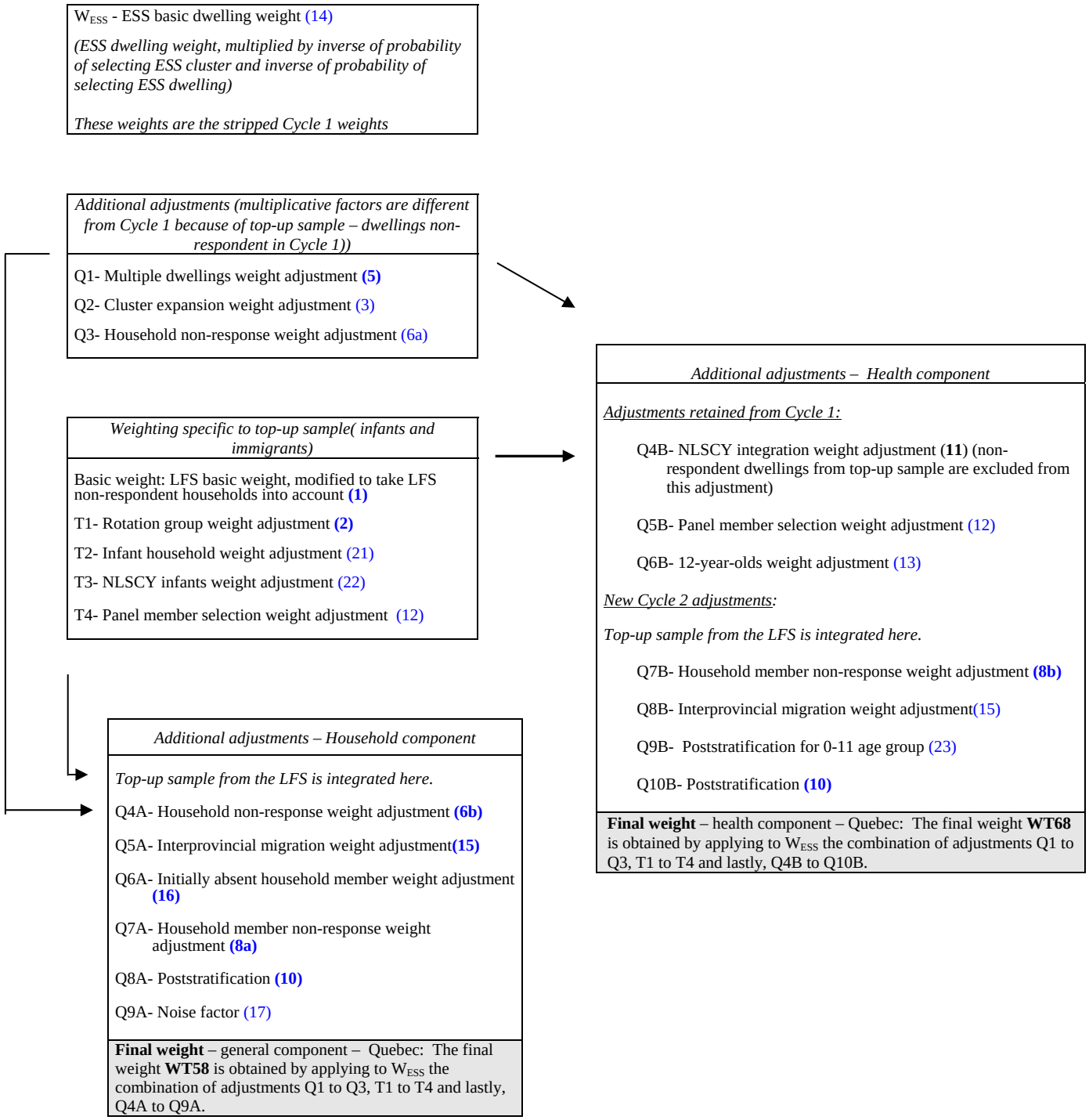


Figure 10.5: Cross-sectional weighting - Cycle 3 - Quebec



10.1.1 Description of the Different Adjustments Shown in Figures 10.1 to 10.5

(1) LFS basic weights:

The LFS sample design is a multistage, stratified clustered design. Initially, clusters were selected with a probability proportional to size. A cluster weight was obtained by taking the inverse probability of selecting the cluster. (Note: in two cases in Winnipeg and one case in Vancouver, three LFS strata had to be combined before selecting clusters. This was taken into consideration when assigning probabilities of selecting these clusters.)

Then, dwellings were systematically selected within these clusters, yielding a dwelling weight (the inverse probability of selecting the dwelling, given that the cluster was selected). An LFS basic weight was calculated by multiplying the cluster weight by the dwelling weight.

$$\textit{Basic weight } (W_{LFS}) = \textit{ cluster weight} \times \textit{ dwelling weight} \qquad (1)$$

Cycle 3 – LFS top-up sample: To select the top-up sample, it was necessary to know the household composition. Non-respondent households (in the LFS interview (see 4.3.2)) were therefore not considered. To compensate for this reduction in the real size of the sample, LFS basic weights were calculated, and they were then inflated slightly to take account of non-response.

(2) Rotation group weight adjustment:

Because the NPHS sample is small compared to the LFS sample, the NPHS retained only a limited number of rotation groups, in each stratum (see 4.1.2). The basic weights are therefore multiplied by the inverse of the proportion of groups used. (Even if only a portion of a group was used to obtain the desired sample size, the whole group is retained in that case. This will be taken into consideration in the stabilization weight adjustment (see (*4*)). The multiplicative factor is:

$$\frac{\textit{Number of rotation groups in an LFS stratum used for the LFS}}{\textit{Full number of rotation groups in an LFS stratum claimed for the NPHS}} \qquad (2)$$

Cycle 3 – LFS top-up sample: For the top-up sample including infants, two rotation groups were used. The weight was therefore 6/2. For the sample of households with new immigrants, four groups were used; the multiplicative factor was 6/4.

(3) Cluster expansion weight adjustment:

The probability of selecting a cluster is based on the census data available at the time of the most recent redesign of the LFS. However, it is possible for the size of the cluster to increase between the census and the point when it appears on the LFS list. Accordingly, subsampling was done within clusters that underwent significant growth. The correction factor consists of multiplying the weights in those clusters by the inverse of the subsampling ratio. The purpose of this adjustment is to avoid excessive work for interviewers in the field.

All provinces except Quebec: The weights were multiplied by:

$$\frac{\textit{Number of dwellings in cluster}}{\textit{Number of dwellings in cluster that are included in subsample}} \qquad (3)$$

For Quebec: The adjustment made depended on the size of the increase.

- Increase under 15%: no adjustment was made.
- Increase of 15% to 30%: the weights of all dwellings in the cluster were increased by the same percentage.
- Increase of more than 30%: supplementary dwellings were selected from among the additional dwellings. For these new dwellings, the weights described thus far were applied, and they were multiplied by the inverse of the proportion of new dwellings selected from among the supplementary dwellings listed. It should be noted that since the new dwellings were not part of the ESS, it was impossible to give different probabilities of selection to each household according to its composition, because the composition was unknown.

(4) Stabilization weight adjustment:

Like the preceding adjustment, stabilization also served to limit the size of the sample. Stabilization consists of subsampling stabilization sectors. These sectors are groups of clusters or strata that have been expanded. Unlike in the preceding adjustment, the expansion is not due to the increased size of any one cluster, but rather to the general expansion of all clusters or strata that comprise the stabilization sector. It is also at this point that the rotation group was subsampled, with the idea of retaining only a portion of it (see rotation group weight adjustment (*2*). The multiplication factor is:

$$\frac{\textit{Number of dwellings selected in an LFS stabilization sector}}{\textit{Number of dwellings used in a stabilization sector for the NPHS}} \qquad (4)$$

(5) Multiple dwellings weight adjustment:

When an interviewer discovered that a dwelling unit included one or more unlisted dwellings, a single randomly selected dwelling was interviewed. The correction factor for this dwelling is equal to the number of dwellings that the dwelling unit actually contained. The weight is therefore multiplied by:

$$\textit{Number of dwellings that the dwelling unit actually comprises} \qquad (5)$$

Cycle 3: This adjustment had to be recalculated in Cycle 3, since multiple dwellings were found in the top-up sample for Cycle 1 non-respondents.

(6a) Household non-response weight adjustment (non-response in Cycle 1):

Some selected households did not respond to the survey. The weights of respondent households were therefore multiplied by:

$$\frac{\textit{Sum of weights of sampled households}}{\textit{Sum of weights of respondent households}} \qquad \textit{(6a)}$$

All provinces except Quebec: The adjustment was calculated separately at the stratum level for each season (summer or winter), since the non-response rate varies significantly from one season to another. The stratum-season classification was adopted because it was the smallest classification guaranteeing stability (corrections almost always less than or equal to 2.5). In the rare cases where the correction was greater than 2.5, it was performed only at the stratum level.

For Quebec: For dwellings that were covered by the ESS, this adjustment was done for each different combination of stratum and household type, for each quarter. The four categories of households were: one-person household, households with children (under 12 years of age), other households with youths (under 25 years of age) and the remaining households (more than one member with no children or youths). If the multiplication factors thus obtained were greater than 2.5, the household categories were grouped.

For dwellings added as a result of an expansion of the cluster by more than 30%, this adjustment was done separately for each cluster and each collection period.

Dwellings outside the scope of the ESS were divided into two weighting classes per collection period. The first class consisted of demolished, vacant or abandoned dwellings and the second consisted of institutions or businesses.

Cycle 2: It was at this stage that the additional cross-sectional sampling units purchased by some provinces in Cycle 1 were removed.

Cycle 3: For all provinces except Quebec, the adjustment was made separately for Cycle 1 household non-respondents in the top-up sample. Owing to the small size of this sample, adjustments were made at the scale of provinces/census metropolitan areas/urban or rural area. Also, all out-of-scope dwellings were dropped. For Quebec, see above.

(6b) Household non-response weight adjustment (non-response in cycles 2 and 3):

As in the case of adjustment (6a), the multiplicative factor is given by:

$$\frac{\textit{Sum of weights of households in sample}}{\textit{Sum of weights of respondent households}} \qquad \textit{(6b)}$$

Cycle 2 – longitudinal sample: It was at this stage that the additional cross-sectional sampling units purchased by some provinces in Cycle 1 were removed.

The adjustment was made separately for each weighting class. To determine these classes, we looked at the characteristics of respondent and non-respondent households, using data collected in

Cycle 1. Categories were defined using a segmentation algorithm that served to distribute households according to certain characteristics. The improved version of the CHAID (Chi-Square Automatic Interaction Detector) algorithm, available in Knowledge Seeker IV for Windows (Angoss Software, 1995), was used to generate the tree structure.

The weighting classes for the general component were based on the characteristics of the dwelling or household, but also on the personal characteristics of the member of the household selected for inclusion in the panel for the health component. These personal characteristics were important for predicting household non-response since, often, non-response was the same for the person as for the household. For example, if longitudinal respondents could not be traced in 1996-1997, the entire household was non-respondent. Also, if the longitudinal respondent was not available or refused to answer in Cycle 2, interviewers were instructed not to interview the household members for the general component.

Separate weighting classes were created for each province. The variables used are shown in Table 10.2 (note that not every province necessarily used all these variables). The same variables were used for the general component as for the health component. However, the non-response classes were different.

Table 10.2 – Variables used to determine weighting classes

Geographic characteristics	Province, census metropolitan area, urban/rural indicator
Household characteristics	Dwelling type, owner/renter status, family type, household income adequacy, main source of income, non-response flag for income in 1994-1995, presence of children in the household
Personal characteristics	Sex, age, age over 16 indicator, marital status, race, country of birth, age at immigration, restriction of activity flag, main activity/labour force status

This adjustment was made at the stratum level. In cases where an adjustment class had a correction factor greater than 2.5, that class was grouped with others until the factor fell below 2.5.

Cycle 2 – RDD buy-in sample: This adjustment was made within each class. If the correction factor was greater than 2.5, the class was grouped with others until the factor fell below 2.5.

Cycle 3: Non-response at the household level for the top-up sample of non-respondent dwellings in Cycle 1 was already dealt with in the previous adjustment. The top-up sample from the LFS was adjusted separately using basic classes because of the small sample size.

(7) ***Rejective method weight adjustment*:**

Some sampled households with no child or youth (under 25 years of age) were rejected in the last two collection periods (see 4.1.2 for more information on the rejective method). To adjust for rejection, the weights of households without youths or children are multiplied by the inverse of (1 – proportion of rejection) for each cluster. The weights of the households concerned are therefore multiplied by:

$$\frac{Number\ of\ households\ sampled}{Number\ of\ sampled\ households\ that\ were\ not\ rejected} \qquad (7)$$

Note: The rejective method was not used for apartment strata, dwellings identified as high income or remote areas. Furthermore, this method was slightly different for Prince Edward Island, since the adjustment was applied to all four periods in that province.

(8a) ***Household member non-response weight adjustment (general component):***

It was possible for some members of a household to have incomplete responses even if the household itself was respondent. The weight of each household respondent is thus multiplied by:

$$\frac{Sum\ of\ weights\ of\ members\ of\ sampled\ households\ for\ the\ non\text{-}response\ class}{Sum\ of\ weights\ of\ respondents\ in\ sampled\ households\ for\ the\ non\text{-}response\ class} \qquad (8a)$$

Cycle 1: Non-response classes are province-age-sex groups (ages 0-11, 12-24, 25-44, 45-64 and 65 and over).

Cycle 2 – longitudinal sample: The non-response rate attributable to a household member was less than 2%. For this reason, the adjustment was made separately for each province only.

Cycle 2 – RDD buy-in sample: The adjustment was made according to age-sex-sub-provincial region categories.

Cycle 3 – The non-response rate attributable to a household member was less than 2%. For this reason, the adjustment was made separately for each province only. Top-up households were adjusted separately, using basic classes.

(8b) ***Panel member non-response weight adjustment (health component):***

It was possible for some panel members to have incomplete responses even if their household was respondent. The weight of each respondent panel member was thus multiplied by:

$$\frac{Sum\ of\ weights\ of\ selected\ persons\ for\ the\ non\text{-}response\ classs}{Sum\ of\ weights\ of\ respondents\ for\ the\ non\text{-}response\ class} \qquad (8b)$$

Cycle 1: Non-response classes are province-age-sex groups (ages 12-24, 25-44, 45-64 and 65 and over).

Cycle 2 –longitudinal sample: Weighting classes, different for each province, were determined in the same way as for the household non-response adjustment for the general component – Cycle 2 (see adjustment *(6b)* for further details and for the lists of variables used). However, the non-response classes were not the same.

Cycle 2 – RDD buy-in sample: The adjustment was made according to age-sex-subprovincial region categories.

Cycle 3: Weighting categories were defined using available information concerning persons selected in Cycle 1. For the top-up sample (non-respondents in Cycle 1 and those from the LFS), separate adjustments were made owing to the small size of the samples.

(9) Weighting specific to the RDD buy-in sample in British Columbia:

Before including the RDD-generated buy-in sample in the core sample, weights were calculated for the buy-in units.

The basic weight is given by the inverse probability of selecting a residential telephone line. The adjustments made to it are similar to those made to the LFS basic weights.

General component weight adjustments (all household members):

- **Multiple telephone lines weight adjustment**: Since a household that has more than one telephone line has more chances of being selected, the weight is multiplied by the inverse of the number of lines, for each household.
- **Household non-response weight adjustment**: see adjustment *(6a).*
- **Household member non-response weight adjustment**: see adjustment *(8a)*. Note: the non-response classes are age and sex.

Health component weight adjustments(panel members only):

- **Multiple telephone lines weight adjustment**: see previous paragraph.
- **Household non-response weight adjustment**: see adjustment *(6a).*
- **Panel member selection weight adjustment**: see adjustment *(12).*
- **12-year-olds weight adjustment**: see adjustment *(13).*
- **Household member non-response weight adjustment:** see previous paragraph.

After making these specific adjustments to the RDD buy-in sample, the sample was integrated with the longitudinal sample. The region covered by the buy-in sample, Prince George in British Columbia, was also part of the core sample obtained from the LFS. The three strata of that province covering the Prince George area were therefore sampled twice. To take this into account, three adjustments to the weights were made for British Columbia as a whole. The first adjustment applied only to the sample obtained by RDD in the three Prince George strata; the second applied to the LFS sample for these same three strata; and the third adjustment concerned the sample obtained from the LFS for British Columbia as a whole, excluding the three Prince George strata. These adjustment factors are fairly complex and are explained in Skinner and Rao (1996).

(10) Poststratification:

Lastly, a poststratification was conducted using population projections based on the most recent census data, adjusted for data from birth and death registers and migration estimates. This poststratification was done separately for each province/age/sex group. Proceeding this way ensured that the sum of the weights within each province-age-sex group would be consistent with the population projections for each of these groups. Since the data were collected over four periods, the population projections used were average projections for these four periods. The multiplicative factor is given by:

$$\frac{\textit{Population projections for the poststratification class}}{\textit{Sum of the weights of respondents in the class}} \qquad (10)$$

For Quebec: The population projections were altered to reflect the fact that the three northern health regions were excluded from the survey.

Cycle 2: The projections were for the 1996 population. For provinces that bought additional sampling units, poststratification was done by health region-age-sex. For the other provinces, it was done by province-age-sex. (Note: households made up entirely of immigrants who arrived in Canada in 1994 or thereafter were not covered by the Cycle 2 longitudinal sample but were included in population projections. Therefore, these immigrants were treated implicitly as if their characteristics were similar to those of the rest of the population.)

Cycle 3: The projections were for the 1998 population. Note that households entirely made up of immigrants who arrived in Canada in 1994 or thereafter were now covered by the cross-sectional survey, owing to the top-up survey. They were also included in population projections.

(11) NLSCY integration weight adjustment:

A member aged 12 and over of each respondent household was selected to become a member of the panel in Cycle 1. In some households, one or more children under 12 years of age instead were selected and administered the NLSCY survey. For more information on the sample design, see Section 4.

These households with children were thus selected, but they were not included in the final sample for the health component of the NPHS. It was therefore necessary to give more weight to respondents from households with children that were part of the NPHS sample. To do this, the weight of panel members from a household with children that was drawn from the adult survey was multiplied by the inverse of the proportion of the total sample that came from the adult sample. The weights of the respondents concerned were therefore multiplied by:

$$\frac{\textit{Number of respondents in panel}}{\textit{Number of respondents in panel who came from adult sample}} \qquad (11)$$

Note: For persons over 12 years of age, the adjustment was done separately for each cluster. For 12-year-olds, it was done separately for each LFS stratum (to be consistent with adjustment *(13)*, described below).

Cycles 2 and 3: Although the NPHS was completely independent of the NLSCY in Cycle 2, the weighting of this cycle was based on the selection of the sample in Cycle 1. It was therefore necessary to account for the presence of the NLSCY in Cycle 1 and apply the NLSCY integration weight adjustment.

Cycle 3: Non-responding Cycle 1 households from the top-up sample were excluded from this adjustment.

(12) ***Panel member selection weight adjustment:***

Since one member aged 12 and over was selected from each household, the probability of selecting each member depended on the number of persons aged 12 and over in the household. The weight adjustment is thus given by the inverse of the probability of selection:

$$\textit{Number of persons aged 12 and over in household} \qquad (12)$$

Note: There was a problem with the CAI application in the first two collection periods of Cycle 1, with the result that no 12-year-olds were selected. An adjustment had to be made in the final two collection periods to compensate for this. This adjustment consisted of giving a greater probability of selection to 12-year-olds than to persons aged 13 and over in the household. In Prince Edward Island, the probability of selecting 12-year-olds was twice as great, while for the rest of Canada it was 1.75 times greater. The adjustment factor was therefore changed accordingly.

Cycle 2 – RDD buy-in sample: For the Alberta and Manitoba samples, both a 12-year-old and over and a child under 12 were selected. For persons aged 12 and over, this means that the adjustment is equal to the number of persons aged 12 and over in the household, while for children, it is equal to the number of children in the household. In Ontario, no child was selected.

(13) ***12-year-olds weight adjustment:***

Because of the problem with the CAI application in the first two collection periods, no 12-year-olds were selected. In order for them to be correctly represented, their weights had to be increased.

All provinces except Quebec: This adjustment was done for each LFS stratum. For households with children, 12-year-olds could have come from the adult sample, regardless of the period, but in reality they were selected only in the last two periods. Since some 40% of the adult sample was interviewed during the last two periods, the weight of 12-year-olds selected during this time was increased by multiplying it by the inverse of the rate, that is, 2.5. Similarly, in households with youths but no children, 12-year-olds could be selected from both the "Adult" sample and the "Child" sample. In the first two periods, the selection was not carried out in the "Adult" sample, unlike what should have been done. Thus, in households with youths but no children, the weights of 12-year-olds were multiplied by the ratio of the percentage of the total sample within an NPHS stratum where they should have been selected to the percentage of the total sample where they were actually selected, that is, by 1.6. Finally, in households with no youths or children, there were no 12-year-olds, so no adjustment was needed. Note that the rates differ slightly for Prince Edward Island, the apartment stratum, dwellings identified as high-income and remote areas.

For Quebec: Children's weights were multiplied by the inverse probability, for 12-year-olds, of being selected in dwellings where a person aged 12 and over was to be included in the longitudinal sample.

(14) ESS basic dwelling weight:

The sample design used for Quebec is based on the ESS sample design (see 4.1.3). The basic weight for dwellings is the ESS dwelling weight, multiplied by the inverse of the probability of retaining the ESS cluster for the NPHS and by the inverse of the probability of retaining the ESS dwelling (this probability includes the probability that the dwelling was retained for the NPHS at the outset and the probability that it was retained because of the composition of the household).

(15) Interprovincial migration weight adjustment:

Sometimes an adjustment was made to the weights for households that moved from a highly populous province to a sparsely populated province. The weights of observations from the more populous province might well be unusually large compared to those for the province of destination. An adjustment was made only if an extreme weight was discovered in an interprovincial migration pattern (leaving one particular province to go to another particular province). In such a case, we looked at the sum of the weights of respondents whose migration involved the same province of departure and the same province of destination. We compared this sum with population projections covering the number of persons who moved in the past two years (for the appropriate provinces of departure and destination). If the sum of the weights was greater than the projections, the weights were reduced so that their sum would be equal to the projections (in general, this is what happened, since the weight that was causing a problem was unusually high). If the sum of the weights was less than the projections, nothing was done; otherwise, we would merely have further increased the extreme weight.

Cycle 3: This adjustment was also made to some individuals who had not moved but had extreme weights.

(16) Initially absent cohabitants weight adjustment:

Only panel respondents were followed from one cycle to another. If the household composition had changed, new household members were included in the general component but no weight was assigned to them since they were not included in Cycle 1. The weight share method (Lavallée, 1995) was a means to solve this problem and assign weights to these new members. It consisted of assigning to each new member of the household the weight of the panel member, divided by the number of members who were in the scope of the survey in Cycle 1 (for example, excluding persons born in 1995 or thereafter and those who had immigrated to Canada since 1995).

Cycle 3: In Cycle 2, infants and persons who had immigrated to Canada since 1995 were excluded because they were not in the scope of the survey in Cycle 1. However, in Cycle 3 this was no longer necessary, since persons not originally included were taken into account by the LFS top-up survey of immigrants and infants. Furthermore, an error in the data collection application made it impossible to select children aged 0 to 11 as members of the panel (health

component) in the top-up sample of non-respondent dwellings in the first cycle. In these households, each member (including children) was assigned the weight of the panel member, divided by the number of household members over 11 years of age.

(17) Noise factor:

For confidentiality reasons, a noise factor was added to the weights of persons within the same household. This factor followed a uniform distribution and was chosen in such a way that the sum of the weights at the household level remained unchanged.

(18) Basic weight for the random digit dialling sample:

In the second cycle, Ontario, Manitoba and Alberta bought a sizable number of supplemental units, selected by RDD. The three provinces were divided into RDD strata (different from the NPHS strata). Each month, a sample of telephone numbers was selected from the RDD sample frame. The initial monthly weight was given by the inverse probability of selecting a particular telephone number. The monthly weights were converted into overall basic weights by multiplying them by the proportion of the sample size of the stratum for the month to the total sample size of the stratum. If the sampling rates or the sample frame did not vary during the survey, all the weights were equal in the stratum.

(18a) Manitoba subsampling weight adjustment:

In several strata in Manitoba, the total number of residential telephone numbers was especially low. The samples selected in these strata therefore included many inactive numbers. For these strata, approximately half the numbers sampled that did not appear on a list of available residential numbers were dropped. To take this into account, the multiplicative factor for the weights was given by the number of non-residential telephone numbers in the sample for the stratum divided by the number of telephone numbers retained and sent to the regional office. This adjustment was made to the non-residential numbers in the stratum that were sent to regional offices. This factor was almost always equal to 2.

(18b) Weight adjustment for telephone numbers selected twice:

The buy-in samples were selected independently of the longitudinal survey sample, and yet the two types of sample covered the same population. It was therefore possible for a household to be selected twice. To avoid this situation, telephone numbers selected by RDD were compared to those in the database of the longitudinal sample. If the number appeared in both places, it was not sent to the regional office for inclusion in the RDD sample. Therefore the weights were multiplied by the inverse of the proportion of numbers actually sent to regional offices. This adjustment was made separately for each stratum.

(18c) Weight adjustment for households with no telephone:

Households with no telephone were also taken into consideration; otherwise, a bias could have been introduced, since such households generally have particular characteristics. It was determined that households with no telephone could be grouped into five categories:

- Lone-person households – person under 65
- Lone-person households – person 65 or over
- Households with two or more persons, all residents 65 and over
- Lone-parent households
- Other households

The weight of households with a telephone was increased by using provincial percentages of households, drawn from the Household Income, Facilities and Equipment Survey conducted in 1989. Adjustment factors were calculated for each province for these five groups.

(18d) Multiple telephone lines weight adjustment:

Dwellings with more than telephone line had a greater chance of inclusion in the survey. In the interview, respondents were asked how many telephone numbers the household had. The weight of households possessing more than one number was multiplied by the inverse of the number of residential lines possessed by the household.

(18e) Collective dwellings weight adjustment:

Collective dwellings accommodating at least 10 unrelated persons could be part of the longitudinal sample but not part of the buy-in sample obtained by RDD. Since the samples were joined, it was important for the buy-in samples to account for collective dwellings. The 1991 census data on residents of collective dwellings were used to calculate the proportion of the population living in such a dwelling. This proportion was calculated separately for each age-sex-marital status group. The multiplicative factor was (1 + proportion of persons in the age-sex-marital status category in a collective dwelling). It should be noted that only the types of collective dwellings included in the NPHS were considered. For example, prisons, hospitals and military bases were not included.

(19) Weight adjustment for integration of longitudinal sample and buy-in samples:

At this stage, the weights for units in the longitudinal sample were combined with the weights of units selected by RDD. So as not to overestimate the population of these three provinces, we used a method for dual survey frames, which was an adaptation of the method developed by Skinner and Rao (1996). We first determined a factor α ($0 \leq \alpha \leq 1$) to indicate the relative size of each sample. We then multiplied the longitudinal sample weights by α and the buy-in sample weights by $1-\alpha$. The factor α was different for each sub-provincial region. In Ontario there were six, in Manitoba there were two and in Alberta, three.

(20) Integration of longitudinal sample and buy-in samples:

To obtain weight sets WT66 and WT66_N, four series of weights were calculated for members of the panel.

For adults (age 12 and over):

- Weight of the "adult" member of the panel
- LFS weight

For children:

- Weight for children
- Weight for children's health care services

Weight of "adult" member of panel

This weight was used to analyse most of the responses in the health component. The core sample and the sample selected by RDD covered the 12 and over age interval. The simplified version of the Skinner and Rao method was used to combine the two samples. Lastly, a poststratification (*10*) was done to obtain the final weights.

LFS weight

This weight was used to analyse responses to some LFS questions. Alberta decided not to ask these additional questions to members of the buy-in sample. For this weight, the two Alberta samples (core and RDD-based) were not combined. The poststratification for Alberta was done at provincial level and not of the health region, since Alberta's buy-in sample was not used. For Ontario and Manitoba, poststratification was performed separately for each health region. In fact, for all provinces except Alberta, this weight was equal to the weight for the "adult" panel member.

Weight for children

This weight is used to analyse most responses in the health component when the panel respondent is under 12 years of age. Once again, the Skinner and Rao method was applied to combine the core sample and the buy-in samples bought by Manitoba and Alberta (Ontario had no children in its supplemental sample). However, this method was altered slightly because in the core sample, unlike in the supplemental samples, there were no children under 2 years of age. Lastly, a poststratification was performed by sex-health region group for Manitoba and Alberta and by sex-province group for the rest of Canada (including Ontario). It was not necessary to include age in the poststratification classes, since all respondents were children (0 to 11 years of age).

Weight for children's health care services

This weight was used to analyse responses to questions on health care services provided to children. These questions were asked only to children selected in the RDD samples in Manitoba and Alberta. For all other provinces, this weight was nil. A poststratification by sex-health region group was performed.

These four series of weights are grouped to form weights WT66 and WT6_N that are found in the different files produced. Weight WT66 is obtained by grouping the *Standard weight for "adult" members of the panel* and the *Standard weight for children*. Weight WT66_N is obtained by grouping the *HPS weight* and the *Weight for children's health care services*.

Weight WT66 applies to all groups in all the provinces. However, it does not cover the 0 to 11 age group in the same way for each province. In Manitoba and Alberta, in households selected by RDD, a child aged 0 to 11 was selected to respond to the health component (in addition to an adult aged 12 and over). For these provinces, WT66 therefore covers children aged 0 to 11. For the other provinces, the only children included were those selected in Cycle 1 (they were therefore between 2 and 11 years of age). WT66 is the weight to use in the health component files and must be used to analyse most of the variables. For the variables obtained from the HPS and questions on children's health care services, the weight WT66_N is used instead.

(21) Infant households weight adjustment:

Among the two LFS rotation groups used to obtain this part of the top-up sample, approximately half of the households with infants were selected. A higher sampling weight was assigned to households with more than one infant, so as to stabilize the weights for the panel member.

(22) NLSCY infants weight adjustment:

In the second of the two LFS rotation groups (used in the third period), households with infants (less than 12 months old) were divided between the NPHS and NLSCY. A second adjustment near 2 was made for these households to compensate for this subsampling.

(23) Poststratification of the 0 to 11 age group:

Whereas the population aged 0 to 3 was correctly represented in the top-up sample of households including infants, the population aged 4 to 11 was not correctly represented in the panel (health component) because of the error in the application (no child aged 0 to 11 could be selected as a member of the panel in the top-up sample of dwellings that were non-respondent in Cycle 1). To compensate for this under-representation, a poststratification was performed. Weights were adjusted to correspond to the 1998-1999 population projections. This adjustment was made for Canada, separately for each age-sex group, with age being divided into two categories (ages 0 to 3 and 4 to 11).

10.2 Longitudinal Weighting

The weighting of the longitudinal panel is different from the cross-sectional weighting for several reasons:

- The longitudinal weighting must represent the population covered when the sample was selected (i.e., 1994 and not the current year).
- The definition of non-response is not the same in a longitudinal context as in a cross-sectional context. For example, persons selected in Cycle 1 who had died or were living

in a health care institution in Cycle 2 are considered respondents from the longitudinal standpoint.

- The longitudinal weighting must not take into account the buy-in samples bought by some provinces or the top-up samples, since these samples are excluded from the longitudinal panel.

10.2.1 Summary of Sets of Longitudinal Weights Produced for the NPHS

Several sets of longitudinal weights were created for each cycle. Each represents a different subset of members of the 1994 longitudinal panel (see section 9 for a description of the types of files). The list below presents all sets of longitudinal weights produced for the NPHS.

- ***Cycle 1:*** A single set of longitudinal weights was created for Cycle 1 data. The weight is WT64LS, and the 17,276 members of the panel are part of the set. Theoretically, the weight variable needs to be calculated only once and can be used from one cycle to the next with the square files. (Note: this weight was actually created in Cycle 2, since no longitudinal file was produced in Cycle 1.)

- ***Cycle 2:*** In addition to the square file (weight WT64LS), two other files were produced in Cycle 2. A first set of weights, called WT66LF, is used for the full longitudinal file. A second set of weights is used for the partial longitudinal file and is denoted WT66LP.

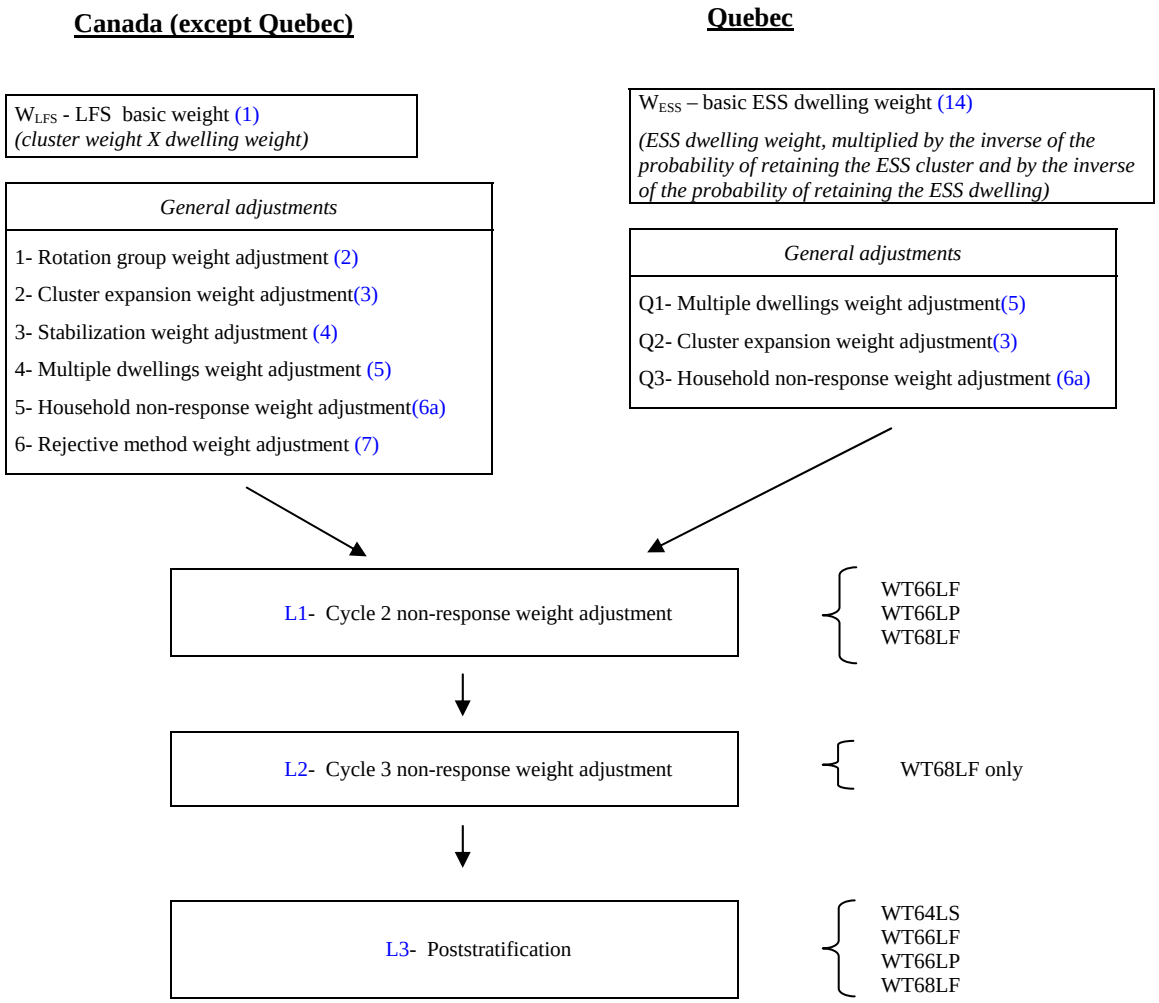
- ***Cycle 3:*** A square file was also produced in Cycle 3 (weight WT64LS). Only one other file was produced: the complete longitudinal file. To create the weight for this file, the starting point was the non-response-adjusted weight for the full file in Cycle 2. An adjustment for non-response to Cycle 3 was applied, followed by a complete poststratification, all for the purpose of creating the weight variable WT68LF.

- *Note regarding share file weights:* As in the case of cross-sectional weighting, an additional adjustment must be made to the weights in the master files in order to obtain weights specific to the share files for each cycle. Because of the small number of respondents who refused to share their responses, this adjustment consisted merely of carrying out a new poststratification, using the same province-age-sex groups as for the initial poststratification (see point *(10)*). It should be noted that for weight WT68_SLF (full longitudinal file in Cycle 3), the variable “Longitudinal response pattern” was also used for this adjustment.

10.2.2 Description of the Different Steps of Longitudinal Weighting

This subsection describes the different adjustments used in longitudinal weighting. Figure 10.6 shows the sequence of adjustments applied specifically to each longitudinal weight. A more detailed description of each weighting step follows the figure.

Figure 10.6: Longitudinal weighting - Cycles 1, 2 and 3



Longitudinal basic weight:

Just as for obtaining cross-sectional weights, longitudinal weights were obtained using the Cycle 1 weights. The basic weights are stripped Cycle 1 weights. These weights are Cycle 1 weights from which some adjustments specific to the Cycle 1 cross-sectional sample have been removed or altered. (For more information, see adjustments 1 to 6 in Figure 10.2 and Q1 to Q3 in Figure 10.3.)

Only two types of adjustment were made to the stripped weights from the first cycle. Most of the adjustments made in order to obtain cross-sectional weights are not necessary, since their purpose was to take into account the change in the probability of selection compared to 1994. New respondents had also been added, which was not the case for the longitudinal panel.

Adjustments to basic weight:

Starting with the stripped weights, adjustments for non-response and a poststratification were performed. The purpose of adjustments for non-response was to redistribute the weight of non-respondent members of the panel, to respondent members. Once again, Knowledge Seeker was used to determine the non-response classes. The classes were defined using the CHAID algorithm.

To correct for panel members who did not respond, the following adjustment was made to the weights of respondent members:

$$\frac{\textit{Sum of weights of all longitudinal members}}{\textit{Sum of weights of longitudinal members responding to the cycle concerned}}$$

Note: A new adjustment for non-response is made in each cycle, and these adjustments are cumulative from cycle to cycle.

(L1) Adjustment for non-response in Cycle 2:

The variables used to determine weighting classes are listed in Table 10.3. (Note: the adjustment for the full and partial sets in Cycle 2 is not the same because the sets are defined differently.)

Table 10.3 – Variables used to determine weighting classes

Geographic characteristics	Province, census metropolitan area, urban/rural area
Household characteristics	Dwelling type, owner/renter status, family type, household income adequacy, main source of income, non-response flag for income in 1994-1995, presence of children in the household
Personal characteristics	Sex, age, age over 16 indicator, marital status, race, country of birth, age at immigration, restriction of activity flag, main activity/labour force status

(L2) Adjustment for non-response in Cycle 3:

In adjusting for non-response in Cycle 3, the characteristics of the household as well as the personal characteristics of the member of the longitudinal panel in 1996-1997 were taken into consideration. As in Cycle 2, a few characteristics related to the survey design or the sampling weight were also taken into account, so as to incorporate the survey design into the analysis. However, unlike the adjustment for non-response in Cycle 2, personal characteristics from the health component were used, since they were available for all records involved in the adjustment for non-response in Cycle 3.

The variables selected by the CHAID algorithm are listed in Table 10.4. The Cycle 1 variable indicating whether the respondent came from the “Adult” sample or the “Child” sample was included, as was a variable indicating non-response to the income questions in Cycle 2.

Table 10.4: Variables used to determine weighting classes

Variable	Description
SEX	Sex
DHC6_AGE	Age
AD_6_1, AD_6_7, ALC6WKY, ALC6_3	Alcohol consumption
AM56_SHA, AM66_SHA, AM66_PXY	Sharing and proxy flags
BPC6_10	Blood pressure
CCC6DNUM, CCC6_1L, CCC6_1N	Chronic conditions
DGC6_1D	Drug use
DHC6_MAR	Marital status
DV_6_65J	Dental visits
EDC6_3	Education
ES_6_80, HCC6F1	Health services use and satisfaction
HSC6DPAD	Restriction of activities
HWS_5	Weight
INC6DIA5, INC6_1A, INC6_3B	Income
INS6_4, INS6_6	Food insecurity
LFC6_41	Labour force
MHC6DWK, MHC6_1A, MHC6_1B, MHC6_1F, MHC6_1L, MHC6_13	Depression
PC_6_40	Had a physical check-up
RPC6_3	Injuries due to repetitive strain
SDC6_4P, SDC6_5A, SDC6_5F, SDC6_6B, SDC6_7A, SDC6_7B, SDC6_7D	Ethnicity
SDC6DAIM	Age at immigration
SMC6_2, SMC6_5, SMS6_9A, SMS6_13A, SMS6_13C, SMS6_13E, SMS6_16D, SMS6_18A, SMS6_18D	Smoking
SHS6_4	Already had sexual intercourse
SP36_CPA	Collection period
SSC6D2, SSC6_3, SSS6_2, SSS6_4	Social commitments and contacts with others

Note: Records for which the selected member was deceased in Cycle 2 or in a health care institution since Cycle 2 are dealt with differently from other records. For these records, no adjustment was made for non-response, since their Cycle 2 weight was already adjusted to take account of the fact that some respondents to Cycle 2 were actually deceased or in a health care institution.

(L3) Poststratification:

Poststratification was used to ensure that the weights for each age-sex-province group summed to the totals for the population, for each group. It is important to keep in mind that the 1994 totals are still used. The longitudinal weights must represent the probability of selection at the time the panel was selected. The adjustment is given by:

$$\frac{\textit{Population estimate in a province-age-sex category}}{\textit{Sum of weights of respondent household members in a province-age-sex category}}$$

11. Data Quality

The evaluation of data quality is an important aspect of any survey. One of the things that it serves to verify is the accuracy or reliability of the information collected. It can also be used to improve the quality of the next run of the survey. At Statistics Canada, this quality evaluation must be designed so as to meet the compulsory minimum requirements of the Policy on Informing Users of Data Quality and Methodology (Statistics Canada, 1998b). These minimum requirements include the production of measures of coverage error, response rates and measures of sampling error. This section mainly covers the measures relating to sampling error. However, a few other measures of quality are examined at the end of the section.

11.1 Sampling Error

Like any sample survey, the NPHS is subject to sampling error. This error depends on a number of factors, such as sample and population sizes, the survey design and the variability of the population. The measurement of this error also depends on the estimation method used. Any Statistics Canada survey must give its data users a means to learn the scope of the sampling error and thus have a better idea of the accuracy of the estimates produced. Two options are offered to NPHS users wishing to obtain a measure of sampling error: calculation of the estimate of sampling error, and use of tables of approximate coefficients of variation (*commonly known as CV tables*). Calculation of the estimate requires much more work, and the calculation technique used has changed considerably over the course of the survey cycles. The next subsection provides more details on this subject. The subsections that follow it deal with the construction and use of CV tables.

11.1.1 Calculating the Estimate of Sampling Error

Since the survey is based on a complex design, and since many adjustments are made to the survey weights to compensate for various factors such as non-response, it is difficult, if not impossible, to calculate variability based on a simple mathematical equation. We will therefore draw on variance estimation methods that are referred to as resampling methods. Carlson (1998) provides a useful overview of the problem as well as the different resampling methods possible. He also examines different computer software that are designed for analysing survey data; more specifically, he looks at the ways in which these software can process data derived from a complex survey design.

While resampling methods are approximate ways of estimating variability, they will be referred to here as being exact methods in comparison with the use of CV tables, which is definitely an approximate method. Over the various cycles of the NPHS, two methods were used: the jackknife and the bootstrap. The history of use of these two methods in the first three cycles of the survey will be reviewed in the next subsection. After that comes a subsection on co-ordinated bootstrap weights, which refers to more recent work in this field, and which seeks to incorporate the dependency that exists between the NPHS cross-sectional samples. Lastly, still with respect to calculating the estimation of sampling error, subsection 11.1.1.3 describes Bootvar, a computational tool supplied to NPHS data users to facilitate the calculation of variance estimates.

11.1.1.1 Estimation Methods Used

Cycle 1

The jackknife method was used for the first cycle of the survey. This method is well-known in the field (see Wolter (1985) for a description of the jackknife method). In the NPHS context, the key point is that using this method requires an acquaintance with the information on the survey design, which in the cases of the NPHS is the definition of strata and clusters. However, since these two variables represent geographic information, they cannot be distributed with public use microdata files because the geographic information that these variables would reveal might affect data confidentiality. This problem situation is discussed in greater detail in Mayda, Mohl and Tambay (1966). Thus the only means available to Cycle 1 PUMF data users to obtain accurate variances/CVs was to have them provided on a cost recovery basis. At the time, the Methodology Division was responsible for responding to these requests. The documentation on the Cycle 1 public use microdata file (Statistics Canada, 1995) provides more information on this subject.

Cycle 2

Different research studies were conducted to solve the confidentiality problem that would result from disseminating the survey design variables. Efforts were made to find a way users could calculate their own estimates of variance and therefore not have to have them provided on a cost recovery basis. Mayda, Mohl and Tambay (1996) describe the different alternatives examined: renumbering of strata identifiers and clusters, creation of pseudo-strata and pseudo-clusters, and collapsing of strata into superstrata. Despite some valuable results, in no case was it possible to guarantee data confidentiality and good estimate quality. The alternative was to use the Remote Access service. This service enables users to e-mail their own programs to Statistics Canada, where the programs are run on files that include the survey design variables, thus making it possible to estimate variances appropriately. The results of the programs are checked to ensure that they do not pose a risk of disclosure, and then they are returned to the users.

Despite these changes, the jackknife method still had a drawback: with the addition of the Cycle 2 buy-in sample, the number of clusters used in the sample had grown considerably, making the jackknife method technically unwieldy. Therefore the search was on for a more effective, more accurate method that could be used by users without having to disseminate confidential information. An examination of the different options available led to the use of the bootstrap method, which had been developed more recently and offered some attractive advantages (see Rao and Wu (1988) for a description of the bootstrap in the general context of a survey with a complex design). For a description of the method and an examination of its main advantages compared to other methods, see the article of Yeo, Mantel and Liu (1999).

In brief, the bootstrap method consists of resampling the total sample many times to create a set of replicates. Each replicate represents a subsample, but it is one in which the survey weights are adjusted to make the subsample represent the total population. The replicate is in fact represented in the final file in the form of a weight variable, commonly called the *bootstrap weight*. The estimation of the variance of a certain parameter may then be obtained simply by calculating the variance of the estimates of this parameter that are obtained with each bootstrap weight generated.

From a technical standpoint, bootstrap greatly reduces the number of replicates needed to make calculations of variance, since unlike jackknife, the number of replicates is not dictated by the number of clusters contained in the sample. Several research studies have looked at the number of replicates to be used. Yeo, Mantel and Liu (1999), on the basis of simulations on NPHS data,

observe that the variance stabilizes when more than 200 replicates are used for estimates of totals and ratios, or when approximately 400 replicates are used for regressions.

Drawing on these findings, the NPHS decided that the number of replicates to be distributed with each survey weight calculated for the survey should be set at 500. The general component of Cycle 2 is an exception to this rule, since the sample of respondents for this component was quite large, and applying the method to it would have entailed using 500 bootstrap replicates. This would have been beyond the capacity of the computers that existed at that time. Therefore the number of replicates for this component was reduced to 100. Lastly, since the cycle 2 and 3 supplemental surveys had much smaller samples, the NPHS produced 2,000 bootstrap replicates for them (see subsection 5.2 for details on these supplemental surveys).

Regarding confidentiality, bootstrap nevertheless provides confidential information, albeit indirectly, so that it is possible to recreate survey design variables (that is, strata and clusters) by comparing the composition of the different bootstrap replicates. Therefore it remained impossible to provide these bootstrap replicates to PUMF users.

Variance estimation is thus done according to the bootstrap method starting in Cycle 2 (and then applied retrospectively to Cycle 1). A file containing bootstrap weights, independent of the data file, is created for each sampling weight. These bootstrap weights are distributed to all users except PUMF users, for whom the estimation of the exact variance must be done via the Remote Access service. Subsections 12.2 and 12.3 provide more details on these different aspects related to variance estimation.

Cycle 3
The situation with respect to variance estimation remained the same as in Cycle 2. Only one small change was made in the creation of bootstrap weights. For cycles 1 and 2, the only adjustment of each bootstrap weight, beyond the basic adjustment for the number of clusters selected in the stratum, was poststratification. For Cycle 3, bootstrap weights were created by incorporating the non-response weight adjustment (this additional stage was carried out only for the health component file because of time constraints during production). Even though a study had shown that adding this stage did not significantly affect variance estimation (see Mantel, Nadon and Yeo, 2000), it was incorporated so that the bootstrap weighting process would more closely resemble the process used for the survey weight.

11.1.1.2 Co-ordinated Bootstrap Weights

Owing to their composition, the cross-sectional samples for the first three cycles of the NPHS cannot be considered to be totally independent. This is because the cross-sectional samples for each cycle include all members of the longitudinal panel that responded to the survey in that cycle, with the rest of the sample coming from buy-ins or top-up samples. Thus, the fact that the cross-sectional samples are partly made up of the same persons implies a degree of dependency between the samples over the cycles, and this dependency must be taken into account when calculating the variance estimate.

The cross-sectional bootstrap weights of the first three cycles were initially calculated without taking this dependency into account. This caused an overestimation of the variance where the statistic calculated involved more than one data cycle, as for example in the case of a test for difference between proportions in cycles 1 and 2.

To remedy this situation, co-ordinated bootstrap weights were calculated. Technically, co-ordinating bootstrap weights means preserving the structure of the bootstrap replicates from one cycle to another for persons that the samples have in common. The bootstrap variance calculation method remains the same, but the use of co-ordinated bootstrap weights serves to take adequate account of the dependence between the samples.

Regarding the calculation of these co-ordinated bootstrap weights as such, it should be noted that unlike what was originally done for cycles 1 and 2, non-response weight adjustments were incorporated into the method. This improvement was first introduced in Cycle 3 (see subsection on Cycle 3 in 11.1.1.1), and it was thus repeated for the preceding two cycles.

Accordingly, co-ordinated bootstrap weights were created for the general component and health files for the first three cycles, and they replaced the existing sets of bootstrap weights.

11.1.1.3 A Computational Tool

Bootstrap weights are distributed to users, and the latter use various documents and programs to facilitate variance calculation using these weights. It should be kept in mind that the situation was different in Cycle 1 because of the use of the jackknife method; since the exact calculation of variance was then done upon a request being made to Statistics Canada, the documentation provided was different. However, to facilitate use of Cycle 1 data, *bootstrap weight files* were created for that cycle, some time after the bootstrap method was adopted. Thus, regardless of the data cycle used, users now have access to the *bootstrap weight file*.

A SAS program is also supplied to users for calculating the variance estimate. This program, called *Bootvar*, consists of several macros which users can activate merely by specifying the names of the variables for which they want to calculate the variance. Although it does not lend itself to the use of all possible analytical methods, Bootvar supports variance calculation for a wide range of statistics: totals, ratios, differences between ratios, regressions (linear and logistic), and general linear models. Complete documentation (in both official languages) accompanies the program and provides examples of its use. A beta version of Bootvar was also developed in SPSS, and during the first three NPHS cycles it was made available only on request. An improved version of Bootvar (version 2.0 in SAS and SPSS) was eventually produced and was disseminated starting in NPHS Cycle 4. This version is much more user-friendly and can be used with data from all cycles of the NPHS.

As noted above, bootstrap weight files may reveal confidential information. To get around this problem, users of public use microdata files wishing to estimate variance using bootstrap must go through the Remote Access service. Details on this service are given in subsection 12.3.

11.1.2 Approximate Coefficients of Variation Tables

The approximate coefficients of variation tables enable users to easily obtain the coefficient of variation (CV) for any estimate of an aggregate (total number of persons), a percentage (of the total population), a difference of aggregates or percentages, a ratio or a difference of ratios. For each of its first three cycles, the NPHS produced a series of tables which are published with the PUMF documentation, and which users of other NPHS products can obtain on request. This

series of tables contains a table for the total population of Canada, one for each province and one for each of the five standard age groups utilized by the NPHS (0-11, 12-24, 25-44, 45-64, 65 and over). Furthermore, these tables are produced independently for each of the two components of the questionnaire. Note that they are valid only for estimates based on cross-sectional data.

Construction of the CV tables entails three steps: i) calculate a number of design effects for the population covered by the table; ii) determine a design effect that is representative of all design effects; and iii) generate the table relating the CV value to the estimated value, based on the design effect selected. Each of these three steps is explained in greater detail below.

i. Calculate design effects

The design effect is defined as the ratio between the estimated variance based on the design effect used and the variance that would have been obtained using a simple random sample of the same size. It serves to give an idea of the efficiency of the sample design and the estimation method.

Since we want to produce approximate coefficients of variation tables that can be used for the variables available in the data file, it will be necessary to determine a general design effect that will be representative of all possible estimates. To do this, the design effect for a mass of variables is calculated for the geography/domain covered by the table in question. The variables used are determined in conjunction with the client division, and they actually represent the variable most likely to be analysed and thus to be used with the table. This list of variables varied from one cycle to another, mainly owing to changes in the questionnaire content.

Note that to generate a larger number of design effects, the design effect of a variable is calculated not only for the total population of the geography/domain covered, but also for several subpopulations.

The variance observed according to the survey design, which is needed to derive the design effect, is calculated for a given cycle using the resampling method utilized in that cycle, that is, with jackknife for Cycle 1 and then with bootstrap for the other cycles.

ii. Determine representative design effect

To then go from the mass of design effects to the one used to construct the table, we take the 75^{th} percentile of the design effects calculated. Choosing this percentile means that the CV table constructed will be more conservative in assigning the CV. When such a design effect is used, in 75% of cases the CV derived from the table will be higher than in reality.

It should also be noted that these 75^{th} percentiles are reported in the PUMF documentation. They may be used to manually make an approximate correction of the variance estimates produced by commercial software, typically calculated on the assumption of a simple random design. When the variances supplied by the software are multiplied by the 75^{th} percentile of the design effect, the user applies a sort of “average” correction to the estimates so as to take account of the design effect of the survey.

iii. Generate tables

Once the "conservative" design effect is obtained, the approximate CV table is generated using the following equation:

$$CV = \sqrt{\frac{D\,N\,(1-\hat{P})}{n\,\hat{Y}}}$$

where D represents the design effect determined (the 75th percentile), N the size of the population examined, $\hat{P}$ the proportion estimated in the population, n the size of the sample used to produce the estimate, and $\hat{Y}$ the estimate of the numerator of the proportion calculated.

Below are a few important details regarding how these tables are produced.

Cycle 1: No table was created for age group 0-11 in the health component, since all children selected were interviewed by the NLSCY (and their data were not part of the NPHS cross-sectional files for this cycle).

Cycle 2: Because of the buy-in of samples and content for this cycle, several tables were added to the series of standard tables. First, for Ontario, Alberta and Manitoba, tables were constructed for each health region (or in some cases, a grouping of health regions). Then for Manitoba and Alberta, the fact that the selected persons in the additional sample could be between 0 and 2 years of age generated a few extra tables: a table for the 0-11 age group exclusively for Manitoba and Alberta, a table for 2-to-11-year-olds in all other provinces combined, and a table for 2-to-11-year olds for Canada as a whole. Two other tables were added to this series because of the purchase of content by Alberta. A few questions in this buy-in content were asked only to respondents in the core sample, and therefore one table was created for the analysis of variables collected from all Alberta respondents, and another for the variables collected only from the core sample. For more information on the approximate coefficients of variation tables for Cycle 2, see the PUMF documentation (Statistics Canada, 1998a).

Cycle 3: Since there was no sample or content buy-in, the tables for this cycle are standard tables, that is, tables for Canada by age group and then for the total population of each province. Note that this was indeed the total population, since that part of the population aged 0 to 4 was covered (cross-sectionally only) by the top-up sample.

11.2 Non-sampling Error

Non-sampling error is associated with many phases of a survey, such as the survey frame, the questionnaire, collection, processing, etc. In short, some form of error in the data can be introduced in almost any step. Even with good control procedures at each phase of the survey, it is impossible to eliminate non-sampling error completely. Not only is this type of error inevitable, it is often hard to measure. The NPHS is not immune to this problem. It produces a few indicators to quantify this error, such as response rates and slippage rates. Response rates are described in Section 8, while slippage rates are discussed in the subsection that follows.

11.2.1 Slippage Rates

Coverage errors occur when sampling units do not adequately represent the survey's target population. One of the indicators used to measure coverage error is the *slippage rate*. By definition, the slippage rate is the percentage difference between the most recent census-based population projections and the population estimates observed for the survey (based on weights obtained before poststratification). The rates generally observed for the NPHS are positive, which indicates undercoverage of the population in the data collection period.

Table 11.1 shows the slippage rates observed for each of the first three NPHS cycles, for each component of the cross-sectional sample as well as for the longitudinal sample. Rates by province are given for the total population, whereas for Canada, rates by age-sex group are included. Note that the slippage rate for the longitudinal sample is the one calculated according to the Cycle 3 square file weight. Theoretically, provincial and Canada-wide slippage rates (longitudinal) should always be the same from one cycle to the next, since the weight adjustment from one cycle to another has always been done on the basis of the person's province of residence in 1994, which was based on the 1991 Census. However, the population projections used for the panel changed some time after the 1996 Census to take the revised projections for 1994 into account. The rates calculated with the cycle 1 and 2 weights are therefore somewhat different from those presented, but they nevertheless remain within the same order of magnitude. As for the rates for Canada by age group, they fluctuate from one cycle to another, since the non-response adjustments that are applied to each cycle do not necessarily take age group into account. Here again, the differences, while larger, are still fairly minimal.

Finally, these explanations also apply when comparing the longitudinal files of a given cycle; the adjustments for going from the square file to the full file take account of the province, but not of age groups. It should be kept in mind that the population projections used for weighting were revised after the 1996 Census, which has the effect of producing different longitudinal weights for the square file for cycles 1 and 2 and the one for Cycle 3, even though these weights should all be similar. Of course, this results in a difference in the slippage rates calculated using the square file for the different cycles.

Table 11.1 Different slippage rates in the NPHS

Region	Sex-age		Cycle 1		Cycle 2		Cycle 3		Longitudinal*
			General	Health	General	Health	General	Health	
Canada	All		10.6	10.7	9.0	9.8	13.0	14.8	10.6
	Males	0-11	3.8	-	6.6	14.7	11.0	25.3	15.5
		12-24	10.8	11.6	6.7	7.4	10.3	14.7	12.5
		25-44	15.9	16.4	14.9	15.2	19.6	20.1	15.8
		45-64	12.8	12.4	8.5	8.4	13.9	12.8	11.1
		65+	6.7	10.7	7.3	8.2	8.1	10.8	9.5
	Females	0-11	7.4	-	7.1	14.6	11.6	26.9	15.2
		12-24	11.4	9.2	8.1	4.2	13.5	9.9	8.1
		25-44	9.9	5.9	9.1	6.3	15.3	10.9	4.4
		45-64	9.3	9.7	7.0	8.8	9.7	9.6	8.6
		65+	10.6	7.6	8.5	10.2	6.3	6.7	4.8
Nfld.	All		6.0	5.6	9.8	9.5	10.7	12.2	6.9
PEI			8.4	8.4	10.1	10.7	11.1	13.1	8.0
NS			8.2	9.0	11.6	11.8	10.9	12.1	8.2
NB			10.9	10.8	8.6	8.3	8.3	9.1	8.4
Qc			9.2	8.4	11.1	11.0	12.2	13.6	9.3
Ont.			9.9	10.4	4.4	6.3	10.6	13.1	9.1
Man.			10.3	11.0	7.7	7.1	9.5	11.7	10.6
Sask.			10.3	11.3	13.1	12.8	13.5	14.9	10.9
Alb.			10.7	11.2	5.8	5.6	15.6	16.5	11.0
BC			16.5	16.6	19.3	20.6	22.1	23.6	18.6

* Obtained from square file for Cycle 3

11.2.2 Other Measures of Non-sampling Error

As noted above, it is often difficult if not impossible to measure some types of non-sampling errors. Béland and Bustros (1998) examined NPHS data quality by looking at aspects such as tracing and response and/or processing error. They found that longitudinal inconsistencies detected over time for a given respondent were not corrected in the data files but were instead left intact in order to leave analysts the choice of how to deal with them.

Much later, Tolusso and Brisebois (2003) took a closer look at some quality indicators relating to non-sampling error, specifically in the context of the longitudinal component of the NPHS. Partial non-response, tracing, refusals and inconsistencies over time are examples of aspects dealt with in this document.

The study of non-sampling errors is an ongoing concern for the NPHS, but this work is often made difficult if not impossible because of a lack of information on sources of errors, or because of a lack of compatibility in the information collected in different cycles.

12. Confidentiality

Confidentiality is an important concern at Statistics Canada, whether it be with respect to data files disseminated or tabulations published or produced in response to special requests. The NPHS invests considerable energy and effort in respecting confidentiality, and the purpose of the subsections that follow is to describe the many efforts made along these lines.

Subsection 12.1 describes the work done to create public use microdata files. Subsection 12.2 focuses more specifically on confidentiality in the field of variance estimation for the NPHS. Lastly, subsection 12.3 examines how a confidential environment is maintained while enabling users to analyse data via Remote Access.

12.1 Public Use Microdata Files

For each of its first three cycles, the NPHS produced different master files: the general file containing all household members (cross-sectional), the health file containing only selected persons (cross-sectional), and a range of longitudinal files. To enable the public to access these data, public use microdata files (PUMFs) were created. Only cross-sectional files could be disseminated, since it was not possible to create a public longitudinal file that satisfied the rules of confidentiality. Indeed, a feasibility study showed that if a longitudinal PUMF was disseminated along with cross-sectional PUMFs, this would risk revealing additional information that might adversely affect data confidentiality. Matching the longitudinal PUMF with one of the cross-sectional PUMFs would give users a greater level of detail than permitted on these public files. A reduction of variables in the longitudinal PUMF might reduce risks of linkage, but any such reduction would have to be so massive that it would render the file useless for analytical purposes. The article of Béland (1999) offers a more detailed discussion of the problems surrounding dissemination of a longitudinal PUMF within the NPHS framework.

Creating a PUMF requires a number of analyses, and a strategy is required. This strategy must be submitted and approved in advance by Statistics Canada's Microdata Release Committee. Once the strategy is submitted, the analyses examining disclosure risks are conducted, so as to identify actions to be taken to protect the anonymity of survey respondents, and ultimately a final PUMF is created. Before it is released, the results of the analyses must be submitted to the Committee for its approval to ensure that the Microdata Release Policy is followed.

Many analyses are conducted to create and validate the PUMF and ensure its confidentiality. The master file is the starting point, since it contains all the variables collected in the survey, as well as derived variables. Note that direct identifiers such as the respondent's name, address and health insurance number do not appear in the PUMF.

The first aspect usually examined is the level of geography that can be reported on the PUMF. The idea is to make sure to have a sizable sample for each geographic area reported, obviously taking account of the area's population size. In general, this has meant that only provinces and an urban/rural indicator could be published in PUMFs for the NPHS. The only exceptions have been for provinces that bought additional units in order to obtain samples representative at the subprovincial regions level, such as health regions. In these cases, a variable identifying the health region (or a grouping of such regions) appears in the PUMF.
Each variable is then examined in relation to various criteria such as its analytical capacity, sensitivity and level of detail. On this basis, the variable will be eliminated, re-categorized or left as is.

Special interest will also be paid to variables considered sensitive or discriminant; these are often referred to as indirect identifiers, that is, variables that cannot be used to identify a respondent without having some knowledge of him/her. For these variables, an effort will be made to ensure that they do not uniquely identify a respondent, either on their own or in combination with other sensitive variables.

The variable representing the survey weight is also examined in detail. Among other things, it will be necessary to ensure that the order of magnitude of the survey weight cannot be used to recreate, with certainty, geographic subregions not reported in the public file.

Since the NPHS general file contains information on all members of the households responding to the survey, it is necessary to ensure that this family perspective cannot be used to identify anyone by allowing the reconstruction of families. In problem cases, it has been possible to suppress or recode some variables.

Lastly, since the NPHS cross-sectional sample was largely made up of the longitudinal panel of each of the first three cycles, it was necessary to make sure that the PUMFs created in each cycle could not be matched in such a way as to recreate longitudinal records. Such matchings would considerably increase the information on a given person and would threaten the confidentiality of the information collected regarding him or her.

This, in a nutshell, is the work required to create a PUMF. Of course, it is necessary not to report some details of the analyses so as to maintain totally the confidentiality of the published data. Once again, the article of Béland (1999) provides further details concerning the work surrounding the creation of PUMFs for the NPHS.

12.2 Confidentiality Regarding Variance Estimation

Since the NPHS is based on a complex survey design, there is no simple formula for estimating variance. It is necessary to turn to approximate methods, such as the resampling methods. One of these, the jackknife method, was used for estimating variance in Cycle 1. This method requires a knowledge of the survey design variables, which are largely geographic in nature. To provide these variables to data users would affect the confidentiality of the PUMF, since it would be possible for a tenacious user to re-derive the survey design variables defining the jackknife subsamples and thus gain access to a few additional variables not available in the PUMF. These variables are especially at risk since they represent relatively small geographic areas, and they would therefore enable users to acquire even more detailed geographic information than allowed in the PUMF and thus improve their chances of identifying particular respondents. For this reason, variance estimation in Cycle 1 was available only on requesting it from the Health Statistics Division.

For Cycle 2, the NPHS had to cope with a sizable number of clusters, which made it very cumbersome to estimate variance using the jackknife method. As described in subsection 11.1.1.1, different alternatives were examined and in the end, the bootstrap method was selected. Because with this method, the number of subsamples to produce was not dictated by the scope of the survey design, the process of estimating variance with the NPHS data was easier.
In shifting to the bootstrap method, we hoped, at the same time, to be able to disseminate bootstrap subsamples to data users so that they could calculate their variance estimates themselves. However, despite the fact that it is even more laborious than with the jackknife

method, it is also possible with bootstrap to recreate survey design variables that define bootstrap subsamples. A variant of the standard bootstrap method, the mean bootstrap (Yung, 1997), was also examined with a view to eliminating this risk of reconstitution, but without success.

Despite the efficiency gain with this new method, the bootstrap subsamples still could not always be disseminated to PUMF users. These users would therefore have to continue to obtain their variance estimates by making special requests to the Health Statistics Division. In light of the growing demand stemming from the popularity of NPHS data, this burden, both on users and on the Health Statistics Division, had to be reduced. The desired solution lay in the Remote Access service recently introduced in some divisions of Statistics Canada.

12.3 Access to Confidential Data via the Remote Access Service

The Remote Access service originated in the Special Surveys Division of Statistics Canada and was then adopted by several other divisions in the Agency. The main goal of this service was to give users access to data that could not be disseminated in a PUMF.

The process enables users, after approval of their proposal, to e-mail their computer programs to Statistics Canada, where the program is run, accessing confidential data files (master files) on Statistics Canada's protected internal network. The computer outputs produced by the program are then vetted to ensure that they do not reveal confidential information, after which they are e-mailed back to the user. The Remote Access service thus gives users indirect access to confidential data and thereby allows them to benefit from greater analytical resources than when using only the limited data contained in the PUMFs. Tambay, Goldmann and White (2001) provide more details on the Remote Access service at Statistics Canada.

To enable users to develop and test their programs on their computer, the NPHS provides them with *dummy files*. These dummy data files replicate the exact structure of the master files used. Creating such a file is an arduous task. Even though the data that it contains are fictitious, the file also attempts to replicate as well as possible the internal consistency that exists between the variables, so that users can not only test the syntax of their programs but also check the sample available for the modelling being considered. Mantel and Nadon (1999) describe in more detail the reasons for creating a dummy file in the NPHS context and the steps involved.

In light of the growing number of dummy files produced in the Agency, guidelines on creating them have more recently been developed. Surveys producing fictitious data must therefore ensure that their products comply with these guidelines, or at least make sure that any deviation from them does not increase the risk of disclosure of information that might put the confidentiality of survey respondents at risk. In fact, since these guidelines were introduced, the strategy for creating a dummy file has had to be formally submitted (usually in writing) to the Microdata Release Committee. Since the dummy files for the first three cycles of the NPHS had all been released when the guidelines were developed, the strategy has never been submitted for these three cycles.

Finally, as noted above, the introduction of the Remote Access service has made the calculation of variance estimates accessible to all users, even PUMF users. Like data files, dummy files of bootstrap weights (bootstrap subsamples) are provided to any user wishing to calculate variance, either with the user's own program or with the Bootvar program, which is also supplied to users (see 11.1.1.3 for more information on Bootvar).

13. Data Dissemination

Like many Statistics Canada surveys, the NPHS must meet the challenge of disseminating as much data as possible while respecting the confidentiality of the information. Table 9.1 shows all the files created for each of the cycles.

For confidentiality reasons, access to the master files is restricted to Statistics Canada employees. However, the general public can access data from these files using different tools developed by Statistics Canada. For example, the Remote Access service, described in 12.3, enables more people to make better and greater use of the data collected in surveys.

Another service offered is cost recoverable client services. Users can submit special requests, such as to have different tables produced. Client Services then performs the work. After verification, results considered not confidential are sent to the users. Another alternative for accessing master files is to come work at Head Office or one of Statistics Canada's regional offices. In this case, users must take an oath and submit to the same rules as Statistics Canada employees. Accordingly they undertake not to disclose any confidential information, subject to penalties. Lastly, in conjunction with various universities in Canada, Statistics Canada has opened Research Data Centres (RDCs), giving interested researchers access to master files. A researcher must first submit a research proposal, which is meticulously studied by a committee, before having access to the data. For further details on the Research Data Centres (RDC) program, visit the website at http://www.statcan.ca/english/rdc/index.htm.

Although master files are accessible only to a limited group of persons under relatively strict conditions, public use microdata files (PUMFs) are accessible to everyone, for a cost. However, it should be noted that academic users can access PUMFs easily and at no charge via the Data Liberation Initiative. For more information on this subject, see the website at http://www.statcan.ca/english/Dli/dli.htm .

Added to this list are dummy files, which are basically master files in which the data have been altered. These files are mainly useful to persons wishing to access master files via Remote Access, having tested their programs in advance. Dummy files of bootstrap weights are also available in the event the user wishes to calculate variance.

Lastly, share files are another means to provide data access to a quite specific clientele. These files contain only the records of respondents who have agreed to share their data with survey partners. The files are thus intended for provincial health ministries, Health Canada, Employment and Immigration Canada (for Cycle 1 only) and Human Resources Development Canada (for Cycle 3 only).

14. Subsequent NPHS Cycles

The need for health information for specific characteristics and for different geographic areas is more critical than ever. This growing demand is coming from policy makers and health care professionals alike, mainly with respect to health care services made necessary by the aging of the population. Health care planners need basic information so that they can better plan for demand and assess the services provided to the population.

The main objective of the NPHS is to measure the health status of Canadians and promote a better understanding of the factors determining health. The main purpose of this survey was to produce not only longitudinal health data based on the responses of members of the panel, but also to produce cross-sectional estimates.

However, even though the NPHS is an excellent source of information, the sample does not contain enough individuals to detect differences in health status for subprovincial regions. Also, because of the sample size, only very limited analysis can be done on special populations such as the elderly, visible minorities, single mothers, users of home care services or persons with specific chronic diseases. To get around this constraint, the Canadian Community Health Survey (CCHS) was introduced in 1999 to produce only cross-sectional estimates. More specifically, the mandate of the CCHS is to fill the main statistical gaps regarding health determinants, health status and the use of the health care system by the Canadian population at the level of health regions.

Consequently, since its fourth cycle, the role of the NPHS has been to focus entirely on longitudinal estimates, leaving it up to the CCHS to produce the desired cross-sectional estimates.

Acknowledgements

The authors first wish to thank all the methodologists who were involved either directly or indirectly in designing the NPHS. None of the authors of this report were among the pioneers of the NPHS. In preparing this report, the authors drew heavily on the studies and documentation left by their predecessors and sometimes had to consult them to obtain additional information. The authors therefore wish to recognize the effort and assistance provided by the original team behind the NPHS. They also wish to thank Sylvain Perron and France Bilocq for their invaluable comments, which served to improve the quality of this report.

Bibliography

ANGOSS Software (1995). Knowledge Seeker IV for Windows – User's Guide. ANGOSS Software International Limited.

Béland, Y. et Bustros, J. (1998). Aperçu global de la qualité de l'Enquête nationale sur la santé de la population (ENSP). Société statistique du Canada - Recueil de la Section des méthodes d'enquête.

Béland, Y. (1999). Release of Public Use Microdata Files for NPHS? Mission ... partially accomplished! American Statistical Association – Proceedings of the Survey Research Methods Section.

Béland, Y. , Bailie, L., Catlin, G. et Singh, M.P. (2000). CCHS and NPHS – An improved Health Survey Program at Statistics Canada. American Statistical Association – Proceedings of the Survey Research Methods Section.

Carlson, B.L. (1998). Software for Sample Survey Data, *Encyclopedia of Biostatistics*, Volume 5, John Wiley & Sons, New York. 4160-4167 (voir aussi article sur Internet à http://www.fas.harvard.edu/~stats/survey-soft/blc_eob.html)

Catlin, G. et Will, P. (1992). National Population Health Survey: First Highlights. Health Report, Vol. 4, No. 3, Catalogue 82-003.

Courtemanche R. et Tarte F. (1987). Plan de sondage de l'Enquête Santé Québec. Cahier Technique 87-02, Montréal, Enquête Santé Québec.

Dolson, D., McClean, K., Morin, J.-P. et Théberge, A. (1987). Sample design for the Health and Activity Limitation Survey. Survey Methodology, Vol.13, No.1, 101-117.

Health and Welfare Canada (1993). Technical Report: Canadian Health Promotion Survey. Published by T. Stephens et G.D. Fowler, Ottawa, Department of Supply and Services Canada.

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. Survey Methodology, Vol.21, No.1, 27-35.

Mantel, H. et Nadon, S. (1999). Dummy File Creation for the Remote Access Program of the National Population Health Survey. SSC Annual Meeting – Proceedings of the Survey Methods Section, June 1999.

Mantel, H., Nadon, S. et Yeo, D. (2000). Effect of Nonresponse Adjustments on Variance Estimates for The National Population Health Survey. American Statistical Association – Proceedings of the Survey Research Methods Section.

Mayda, J.E., Mohl, C., et Tambay, J.-L. (1996). Variance estimation and confidentiality: They are related. Proceedings of the Survey Methods Section, Statistical Society of Canada, 135-141.

Mohl, C. (1990). National Population Health Survey Institutional Sample. Statistique Canada, Division des méthodes d'enquêtes auprès des ménages, document interne.

Norris, D. et Paton, D. (1991). Canada's General Social Survey : Five Years of Experience. Survey Methodology, Vol. 17, No. 2, 245-260.

Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. Journal of the American Statistical Association, 83, 231-241.

Singh, M.P., Drew, J.D., Gambino, J. G et Mayda, F. (1990). Methodology of the Canadian Labour Force Survey 1984-1990. Statistics Canada, Catalogue n° 71-526.

Singh, M.P., Gambino, J. et Laniel, N. (1994). Research Studies for the Labour Force Survey Sample Redesign. American Statistical Association – 1994 Proceedings of the Section on Survey Research Methodology, pages 715-720.

Skinner, C.J. et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex survey designs. Journal of the American Statistical Association, 91, 433, 349-356.

Statistics Canada (1995). National Population Health Survey, 1994-95: Public Use Microdata Files – Household Component .Statistics Canada, Catalogue N° 82F0001XDB.

Statistics Canada (1998a). National Population Health Survey, 1996-1997: Public Use Microdata Files – Household Component. Statistics Canada, Catalogue N° 82M0009GPF.

Statistics Canada (1998b). Statistics Canada – Quality Guidelines: Thrid Edition – Octobre 1998. Statistics Canada, Catalogue N° 12-539-XIF.

Statistics Canada (1998c). Methodology of the Canadian Labour Force Survey. Statistics Canada, Catalogue no. 71-526-XPB.

Statistics Canada (2000). National Population Health Survey, 1998-1999 (Cycle 3): Public Use Microdata Files – Household Component. Statistics Canada, Catalogue N° 82M0009GPF.

Statistics Canada (2002). Population Health Survey Program - National Population Health Survey. Cycle 4 (2000-2001), Household Component. Longitudinal Documentation to go with the Longitudinal Data Set, May 2002.

Statistics Canada (2003a). Microdata User Guide National Longitudinal Survey of Children and Youth, Cycle 4.

Statistics Canada (2003b). http://www.statcan.ca/english/concepts/nphs/nphs1_e.htm.

Stukel, D., Mohl, C. et Tambay, J.-L. (1997). Weighting for Cycle Two of Statistics Canada's National Population Health. Proceedings of the Survey Methods Section, Statistical Society of Canada, 111-116.

Tambay, J.-L., Goldmann, G. et White, P. (2001). Providing Greater Access to Survey Data for Analysis at Statistics Canada. American Statistical Association – 2001 Proceedings of the Section on Survey Research Methodology.

Tambay, J.-L. et Mohl, C. (1995). Improving Sample Representativity through the Use of a Rejective Method, Actes de la conférence de l'ASA, Section on Survey Research Methods, p: 29-39.

Tolusso, S. et Brisebois, F. (2003). NPHS Data Quality: Exploring Non-sampling Errors, Statistics Canada, Document de travail de la direction, HSMD-2003-004E.

Wolter, K.M. (1985). Introduction to Variance Estimation. Springer Series in Statistics. ISBN 0-387-96119-4.

Yeo, D., Mantel, H. et Liu, T.P. (1999). Bootstrap Variance Estimation for the National Population Health Survey, 1999 Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 778-783.

Yeo, D. (1999). After the First Steps: The Evolution of a Longitudinal Survey, Statistique Canada, paper presented at the Workshop on Longitudinal Research in Social Science – A Canadian Focus, Population Studies Centre, University of Western Ontario.

Yung, W. (1997). Variance Estimation for Public Use Microdata Files. Proceedings of Statistics Canada Symposium 97 - New Directions in Surveys and Censuses, Statistics Canada, No. 11-522-XPE.