

# CHALLENGES IN DEVELOPING THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN

Marie Brodeur, Gilles Montigny, H el ene B erard, Statistics Canada  
Marie Brodeur, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6

**Key Words: Household, Frame, Sample, Weighting**

## Introduction

Statistics Canada, in co-operation with the federal department of Human Resources Development Canada, has developed the National Longitudinal Survey of Children (NLSC). The purpose of this survey is to gather information for policy analysis and program development on critical factors affecting the development of children in Canada.

The first cycle of data collection began in November 1994 and extended to June 1995. The initial sample for this first wave included some 29,000 children from 0 to 11 years of age. Data collection will be repeated at two-year intervals. The children originally surveyed in the first wave will be followed over time, presumably until adulthood. Also, the sample for subsequent collection will be augmented to represent children in age groups no longer covered by the longitudinal sample.

Participating households (individuals occupying a dwelling) in the initial wave of the NLSC were selected from the Statistics Canada's Labour Force Survey (LFS) frame. The Core component of the NLSC sample (25,000 children) was drawn directly from rotation groups of the LFS frame while the remaining 4,000 children came from an integrated component with the National Population Health Survey, another longitudinal survey initiated by Statistics Canada in 1994.

This report describes the methodology used in the NLSC. The first section presents the objectives of the project. The NLSC longitudinal concept is described in section 2. The third section presents the sampling plan for the first cycle. The fourth section focuses on the different collection methods and introduces the various components of the questionnaire. The last section presents an overview of the weighting and estimation methods.

## 1. Objectives

The primary objective of the NLSC is to develop a national database on the characteristics and life experiences of Canadian children as they grow from

infancy to adulthood. Initially, its focus is on young children. More specifically, the main objectives of the NLSC are:

1. to determine the prevalence of various biological, social and economic characteristics and risk factors in Canadian children and youth;
2. to monitor the impact of such factors, life events and protective factors (including interventions) on these children's development;
3. to provide this information to policy and program officials for use in developing effective policies and strategies to help children to live healthy, active and rewarding lives.

Underlying these objectives is a need to fill an existing information gap describing a broad range of characteristics of children in Canada, in particular covering the early developmental years. As such, a holistic approach to measuring child conditions and characteristics is essential to capture, to the extent possible, the diversity and dynamics of factors affecting the development of children into self-reliant adults. The focus of this study is on environmental factors such as the family, the school and the community as well as measuring various outcome indicators throughout the child's development years.

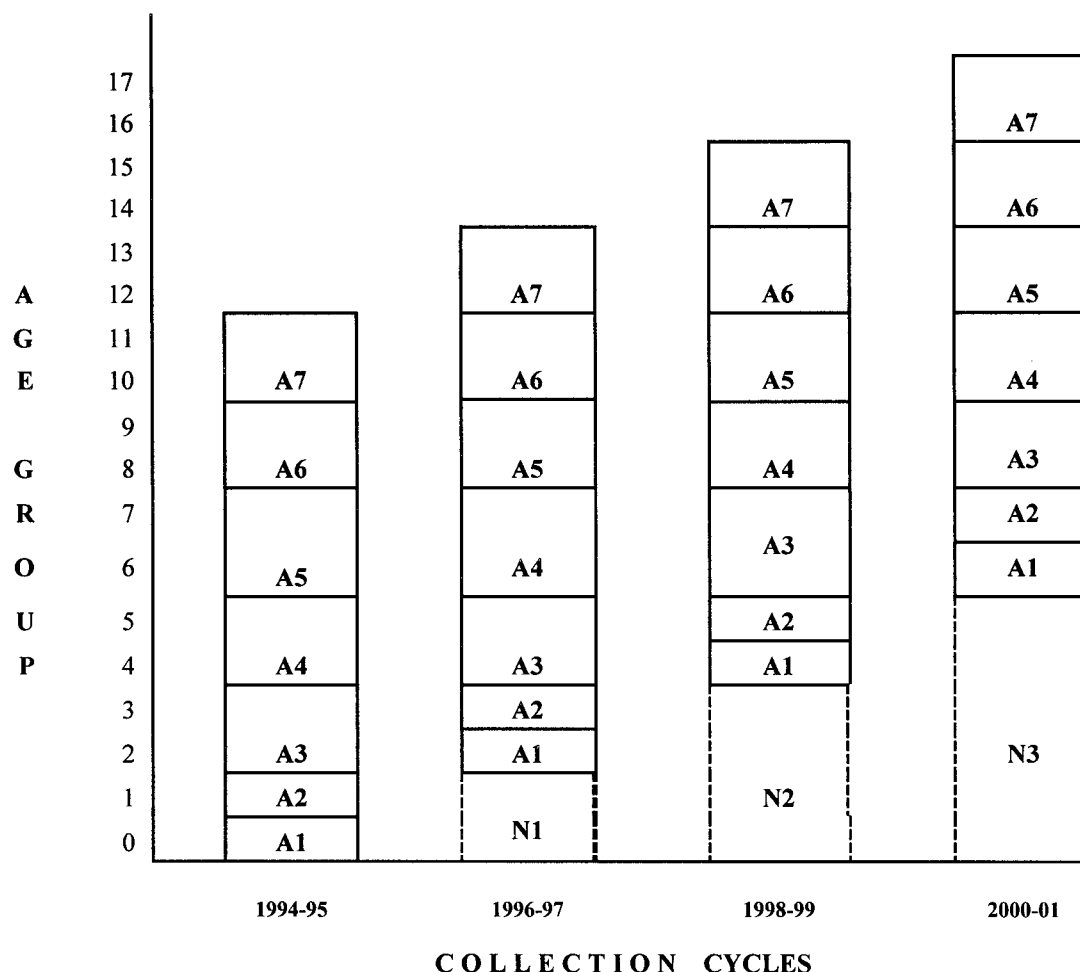
## 2. Longitudinal concept

The design of the NLSC consists of a sample of children, selected in 1994, to be followed over time. The sample consists of an original longitudinal sample, augmented to meet periodic cross-sectional data needs. Figure 1 illustrates the first four collection cycles of the NLSC.

For the first collection cycle, the survey yields cross-sectional information on the population aged 0-11. The grouping of the sample into newborn, children aged 1, 2-3, 4-5, 6-7, 8-9 and 10-11 was adopted to allow for two year age spans between collection cycles. This grouping facilitates analysis every two years, while maintaining

**Figure 1: NLSC Sample Design**

(through four survey cycles)



emphasis in the youngest age groups since newborns and one year olds are retained as separate groups.

For the longitudinal component of this survey, any child aged 0-11 for whom information is collected in the first year remains in the sample for the duration of the survey. This means that if the child moves out of the sample dwelling at some later date, that child is traced, and the next wave of interviewing is conducted at his or her new location. From a longitudinal perspective, the child, not the dwelling, is the unit of sampling.

For the collections following the initial collection, the current plan is to upgrade the sample in age groups no longer covered by the longitudinal sample in order to maintain coverage of the lower age ranges (see N1, N2 and N3 on the chart). This sample would be selected from

LFS rotation groups. For example, for the second survey cycle, N1 new children (ages 0,1) would be added to the original sample.

Current plans are to use the "augmented sample" only for cross-sectional purposes. This allows the production of cross-sectional estimates at least for the early collection cycles. For example, cross-sectional estimates would be produced for children 0-13 years of age in 1996 and for children 0-15 years of age in 1998. The children added to the original sample would not be followed longitudinally.

The National Longitudinal Survey of Youth (NLSY), conducted by the United States Bureau of Labor Statistics, had an attrition rate of 9% (Baker (1993)). The NLSC can be expected to have a similar attrition rate. At the present time, no updates of the sample for attrition are planned.

### 3. Sampling Plan for the First Cycle

The NLSC uses several sampling frames, all coming from the Canadian Labour Force Survey (LFS). The following section describes the NLSC population, presents the sampling frames and shows the sample allocation.

#### 3.1 Definition of the NLSC population

The target population of this survey consists of Canadian children 0 to 11 years of age. It should be noted that nearly a third (28%) of Canadian households have children within this age range. The survey population is a subset of the target population and excludes:

1. children residing in institutions (roughly 0.5 percent of children 0-11 years of age reside in institutions);
2. native children living on reserves.

However, the NLSC survey includes the following:

1. households with a child or children that include civilian employees of the Armed Forces or members of the Armed Forces occupying a dwelling on or off a military base;
2. households with a child or children of aboriginal people not living on reserves.

A separate longitudinal study is conducted in the Yukon and the Northwest Territories.

#### 3.2 Description of the LFS

The LFS is a monthly household survey conducted by Statistics Canada, covering some 55,000 households throughout the country. It is used to produce monthly estimates of employment, self-employment and unemployment. The survey covers approximately 97% of the population of Canada. The population of the Yukon and the Northwest Territories, persons living on Indian reserves, full-time members of Canada's Armed Forces living in barracks and residents of institutions are excluded from the LFS. Civilian employees of the Armed Forces and aboriginal persons living "off reserve" are covered by the survey.

This survey uses a multi-stage sample design based on an area frame in which dwellings (residences) are the sampling units. All eligible individuals who occupy one of the dwellings selected are included in the LFS sample. In other words, the LFS draws a sample of dwellings, and data are collected on the various individuals in the

households. The sample consists of six rotation panels (groups). Each month one of the six rotation groups is replaced after being included in the sample for six months. Each rotation group contains some 10,000 households, representing roughly 25,000 individuals. For more details, see Singh and al. (1990).

A major redesign of the LFS was conducted in 1995 using information from the 1991 Census to update the frame. The stratification methods and sample allocation were also modified. Phase-in of the new design started in October 1994, one rotation group at a time. The new sample was completely in place in March 1995. For the purpose of this text, the LFS frame prior to the redesign is called the Old LFS frame while the LFS frame after the redesign is called the New LFS frame.

#### 3.3 Description of the National Population Health Survey

The National Population Health Survey (NPHS) was going out in the field at about the same time as NLSC, to collect information related to the health of the Canadian population. Most of the information on the household and on children's health required by NPHS is identical to that collected in NLSC. To avoid duplication of effort and to lessen response burden on Canadian households, it was decided that the sample and the content of these two surveys should be partially integrated. This means that in households where the selected person in NPHS is a child under the age of 12, the household receives the NLSC interview.

NPHS uses the LFS frame, with some modifications, to meet some Health criteria by province. It takes a fresh sample of households. These are households that have never been in the LFS. More information regarding the sample design and the selection of households with children is given in Tambay (1994).

#### 3.4 The NLSC Frame

The NLSC frame is composed of three frames drawn from the LFS frame, all covering the same population. The first frame is the Old LFS frame, the second frame is the New LFS frame and the third frame is the NPHS frame. The first two frames are considered independent because the stratification process was completely redone and the two samples were selected independently. There are no Primary Sampling Units (PSU) shared. However, the second and third frames are not independent. They both come from the New LFS frame and they share some

PSUs. This creates some estimation problems, which will be covered in section 5.2.

Approximately 2,500 households from the NPHS frame have been selected for the NLSC interview. Data are collected for the NLSC from up to 4 children under the age of 12 in these selected households, yielding an NLSC sample of approximately 4,000 children aged 0-11. This part of the survey is known as the Integrated component. The remaining sample of 25,000 children is selected from the first and second frame. This problem does not occur in the province of Québec where three separate frames were drawn from the Old LFS frame, the New LFS frame and l'Enquête Sociale et de Santé.

Analysis of data available from the LFS has indicated that approximately 30% of households in the LFS sample have children under the age of 12. The two options considered for the sample design were:

1. A sample of new households from the LFS frame
2. A sample of households currently participating in the LFS.

Due to the relatively low incidence rate of households with young children, it was decided that the use of a new sample of households, for which no previous information is available, would be inefficient. These households would need to be screened to find those with children. The second option was chosen since the composition, the number and the age of children are known in households that were previously interviewed for LFS. Therefore, only households that have children in the appropriate age range are contacted for the NLSC.

### 3.5 Sample allocation

For the NLSC, the sample size had to be sufficiently large to produce reliable estimates in each of the following age groups: 0, 1, 2-3, 4-5, 6-7, 8-9, 10-11 at a national level. (No specific characteristics are being measured, but it is assumed that some health or behaviour problems could affect around 4% of the children (by age group)). The objective is to estimate a minimum proportion of 4% for each group to obtain reliable estimates at the national level. A second requirement is that sufficient sample be selected in each province to produce reliable estimates for children 0-11 (not broken down by age group).

A square-root allocation to provinces was used to allocate the 29,000 children using the average number of children in the LFS. The size of the proportion to be estimated, the required precision of estimates, the design effect

associated with using the LFS design, and the overall response rate after a number of contacts were considered. A design effect of 2.0 is assumed for the LFS design. It was required that the minimum proportion of 4% be estimable with a coefficient of variation (C.V.) of 16.5% or better. The LFS requires that estimates have a C.V. of 16.5% or better to be releasable without restriction. The NLSC is using the same criterion.

After five waves, an overall response rate of 55% is to be expected (80% in the first wave, 91%<sup>1</sup> in subsequent waves). The last requirement was to distribute the 29,000 units between the Integrated and Core components.

The sample was first allocated to the provinces to evaluate how many rotation groups were necessary to meet the objective of 29,000 children. It was found that 9 rotation groups were sufficient.

The number of rotation groups needed for each age range varies by province (see table 1). The sample size by age and province varies due to factors such as the sample allocation of the LFS and the number of children per household in each province. An average number of 9 rotation groups are used except in Ontario and Alberta where 8 rotations are used.

As mentioned previously, the sample allocation was done using the average number of children in the LFS. The number of children varies from one month to another depending on whether the sample is in a neighbourhood with young families or one with grown-up children. Therefore, the sample allocation is an estimate of what could be anticipated. It was possible that fewer or more children would be interviewed.

To reduce the respondent burden within a household and to limit the interview length, a maximum of 4 children age 0-11 were randomly selected within a household. This affects between 1% and 2% of households in Canada.

## 4. Content and collection methods

The content of the NLSC was developed in consultation with child development specialists and with child program administrators within the federal and provincial governments. In order to optimize data collection, content

---

<sup>1</sup> A 9% attrition rate was found by the National Longitudinal Survey of Youth in the US and the Survey of Labour and Income Dynamics (SLID) from Statistics Canada.

**Table 1. Number of children by province and number of rotation groups.**

PROVINCE	Number of children			MIN p	Number of rotation groups
	Integrated	Core	TOTAL		
Newfoundland	264	1,300	1,564	7.9%	9
Prince Edward Island	193	773	966	12.1%	9
Nova-Scotia	232	1,659	1,891	6.6%	9
New Brunswick	255	1,544	4,421	6.9%	9
Quebec	746	4,499	5,245	2.5%	9
Ontario	1,239	6,299	7,538	1.7%	8
Manitoba	355	1,850	2,205	5.7%	9
Saskatchewan	265	2,019	2,284	5.5%	9
Alberta	351	2,330	2,681	4.7%	8
British Columbia	373	2,313	2,686	4.7%	9
<b>TOTAL</b>	<b>4,271</b>	<b>24,586</b>	<b>28,857</b>	<b>0.5%</b>	

**Table 2. Number of children by age groups.**

AGE GROUPS	Integrated	Main	TOTAL	MIN p
0	352	3,051	3,403	3.8%
1	314	2,728	3,042	4.2%
2-3	711	4,092	4,803	2.7%
4-5	713	3,912	4,625	2.8%
6-7	715	3,620	4,335	3.0%
8-9	734	3,616	4,350	3.0%
10-11	734	3,567	4,301	3.0%
<b>TOTAL</b>	<b>4,271</b>	<b>24,586</b>	<b>28,857</b>	<b>0.5%</b>

selection took into consideration criteria such as the importance of the policy or scientific issue to be addressed, the segment of the population covered and the robustness of the measurement tools available for a household collection context. These issues are discussed in detail by McKellar and all (1993). The content and the data collection methodology are presented in the following sections.

#### 4.1 Content

The content is structured into nine components:

1. The household record - contains the household

members' names, their relationships, age, marital status and administrative information.

2. The general component - designed to measure the following topics: restriction of activities, chronic conditions, socio-demographic characteristics, education, labour force activity and income

3. The parent component - covers parent/adults health and social support, family functioning and neighbourhood characteristics.

4. The child component - designed to provide detailed information on subject areas most relevant to each age

group such as:

- health (all)
- family history and custody arrangements (all)
- child care (all)
- medical/biological factors (ages 0 to 3)
- temperament (age 0 to 3)
- behaviour (all)
- motor and social development (ages 0 to 3)
- parenting style (all)
- activities/time use (all)
- literacy activities (all)
- education (ages 4 to 11)
- relationships (ages 4 to 11)

5. A self-administered questionnaire, for children aged 10 and 11 - collects information from children on various subjects such as relationships, school experience, feelings and behaviours, self-esteem and activities.

6. A teacher's questionnaire - gathers information on the children's behaviour, academic performance, class, teaching practices and school climate.

7. A principal's questionnaire - collects information on the school attended by the child, information related to the allocation of students to classes, student and teacher composition within the school and school climate.

8. A mathematics achievement test - administered by the teacher to children in grades 2 and over.

9. A test of receptive vocabulary - given to children 4 and 5 year old (Peabody Picture Vocabulary Test (PPVT)).

#### 4.2 Collection methods

For this survey, several collection methods were used. The main collection was conducted by way of personal interviews at the respondents' homes. In addition, some information was collected with a self-administered questionnaire and some with mailed questionnaires.

For the first cycle, the data were collected over four collection periods, two collection periods for the Integrated component and two collection periods for the Core component. The first and fourth periods were for the Integrated sample and were conducted in November 1994 and March 1995. About 2,000 children were surveyed in each of these two collection periods. The second and third periods were for the collection concentrating on the 25,000 NLSC children drawn from the LFS data base, and they were conducted in December 1994 and in February

1995. Finally, a follow-up was conducted in June 1995.

The household collection was conducted by way of personal interviews at the respondent's home. Personal computers were used to administer the interview. This is known as Computer-Assisted Personal Interviewing (CAPI Component). Data were gathered from parents (or a person knowledgeable about the child). Children 10-11 were asked to complete a self-administered questionnaire. The median length of the interview was 2 hours for a family of two adults with two children.

The teacher of the child and the principal of the school, identified by the household respondents, were each asked to complete a questionnaire. The collection method was essentially as follows: for each teacher/principal identified, a mail-in, mail-out questionnaire, with follow-up, was sent to the child's school address. Data collection from teachers and principals took place after the data collection from households.

#### 5. Weighting

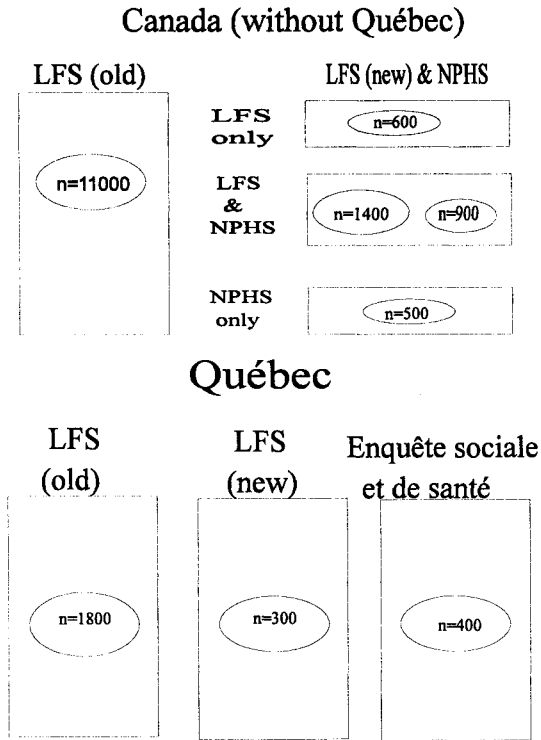
The first year of the survey is essentially a cross-sectional survey. The first step to obtain the cross-sectional weight is to take the inverse of the selection probability corrected for non-response. After this basic step, two approaches could be performed to obtain the weight. This section presents a general overview of the weighting strategy.

##### 5.1 Weighting Approach

Two different approaches could be used to weight household data; individual weighting and integrated weighting. In both cases, weights are adjusted to match geography-age group-sex control totals. In the latter, every member of a household is controlled to have the same weights. This approach is used by the LFS and other special surveys to counteract the following problem noted by Lemaître and Dufour (1987). "Because of the age-sex adjustment factors, the weight so assigned will usually differ from person to person within the same household. When estimating characteristics of persons, this may not pose any particular problem; in producing estimates of households or families, however, it is not entirely clear which weight is the appropriate one to use, if any".

The integrated method is very useful for data analysis. However, some disclosure avoidance problems arise especially for longitudinal surveys. It is easier to identify a household because every individual has the same weight. This issue is very crucial for longitudinal surveys because after two collection waves, it becomes easier to

**Figure 2**  
**NLSC Sampling Frame**



identify respondents. If an integrated weighting approach was selected, the number of variables on a micro-data file would have to be reduced. Furthermore, the concept of a longitudinal household is difficult to implement because household membership can be different in successive waves after death, birth or divorce. The integrated approach would not be relevant to calculate a longitudinal weight.

Other longitudinal surveys at Statistics Canada, like NPHS and SLID are using individual weighting. To harmonize the methodology and to avoid the problems mentioned previously, an individual weight was adopted.

The NLSC data will be released on a micro-data file where there will be a record for every child in the sample. It is not planned to release information at the household level. Therefore, an individual weight will be created for every child. The LFS and NPHS sub-weights (the inverse of selection probability) will be adjusted for non-response. However some theoretical background had to be developed to combine the three samples from the three different frames.

## 5.2 Weighting Strategy

As mentioned in section 3.4, the NLSC sample was selected from three different frames. The next section describes the weighting strategy for all provinces except Québec while the second section will provide the weighting strategy for Québec.

### 5.2.1 All provinces except Québec

The sub-samples to combine are the following :the Old LFS sub-sample ( $n_A$ ), the New LFS sub-sample ( $n_B$ ), the NPHS sub-sample( $n_C$ ) and the combined sub-sample (shared PSUs) of the New LFS frame and the NPHS frame ( $n_{BC}$ ) (see Figure 2). The sample sizes (number of households) of each sub-sample are indicated to show their relative importance. A composite estimator is used to estimate these sub-samples. The formula is given by:

$$\hat{Y} = \alpha \hat{Y}_A + (1 - \alpha) \beta \hat{Y}_B + (1 - \alpha) \gamma \hat{Y}_{BC} + (1 - \alpha) (1 - \beta - \gamma) \hat{Y}_C$$

where  $\hat{Y}_A$ ,  $\hat{Y}_B$ ,  $\hat{Y}_{BC}$  and  $\hat{Y}_C$  are the estimates obtained from the four sub-samples.  $\hat{Y}_{BC}$  is calculated using a weighted sum of the households coming from the shared PSUs. More details are provided in Brodeur and Bérard (1995). The alpha, beta and gamma coefficients are given by:

$$\alpha = \frac{n_A}{n_A + n_B + n_{BC} + n_C}$$

$$\beta = \frac{n_B}{n_B + n_{BC} + n_C}$$

$$\gamma = \frac{n_{BC}}{n_B + n_{BC} + n_C}$$

The sample size proportions were used to determine the coefficients because they remain stable for domain estimation as opposed to an optimal method.

### 5.2.2 Province of Québec

As illustrated in Figure 2, the three sub-samples for the province of Québec are: the Old LFS sub-sample, the new LFS sub-sample and the NPHS sub-sample derived from l'Enquête Sociale et de Santé (see Tambay (1994)). These three sub-samples are independent. A composite estimator is used again and is given by the following formula:

$$\hat{Y}_Q = \alpha \hat{Y}_A + \beta \hat{Y}_B + (1-\alpha-\beta) \hat{Y}_S$$

where  $\hat{Y}_A$ ,  $\hat{Y}_B$  and  $\hat{Y}_S$  are the estimates obtained from the three sub-samples, the New LFS sub-sample and Enquête Sociale et de Santé sub-sample. The coefficients alpha and beta are:

$$\alpha = \frac{n_A}{n_A + n_B + n_S}$$

$$\beta = \frac{n_B}{n_A + n_B + n_S}$$

### 6. Conclusion

The NLSC is an ambitious project. This paper has presented an overview of the first cycle. More work has to be performed on the first cycle and its initial design. The plans are to use the JACKKNIFE method to estimate the variance.

The second cycle will be conducted in 1996. The longitudinal strategy is now being finalized. Follow-up rules of children will need to be formulated and the method to derive a replacement sample for children aged 0 to 1 will be examined. Finally, a strategy for the derivation of longitudinal and cross-sectional weights will need to be elaborated in preparation for the release of the second cycle data.

### Bibliography

Baker, P.C., Keck, C.K., Mott, F.L., Quinlan, S.V. (1993), NLSY Child Handbook. Center for Human Resource Research, Ohio State University, Columbus, Ohio.

Brodeur, M., Bérard, H. (1995), National Longitudinal Survey of Children. Description of the NLSC Methodology. Social Survey Methods Division, Statistics Canada.

Brodeur, M., Bérard, H. (1995), NLSC Weighting and Estimation. Social Survey Methods Division, Statistics Canada.

McKellar, S., Montigny, G., Grainger, P. (1993), National Longitudinal Survey of Children: Description of Content. Statistics Canada Working Paper Catalogue No. 93-06R.

Lemaître, G., Dufour, J. (1987), An integrated Method for Weighting Persons and Families, Survey Methodology, Vol.13, No.2. Statistics Canada.

Singh, M.P., Drew, J.D., Gambino, J.G., Mayda, F. (1990), Methodology of the Canadian Labour Force Survey. Statistics Canada Catalogue No. 71-526.

Clark, C., Drew, D., Paillé, B. (1994), Upcoming Changes to the LFS. Statistics Canada, Internal report.

Tambay, J.L., (1994), Sample design of the National Population Health Survey. Statistics Canada.