Table of Contents

1 INTRODUCTION

The Youth Smoking Survey (YSS) was conducted by Statistics Canada in the fall of 1994, on behalf of Health Canada. The prime objective was to provide information on smoking behaviour of young people in the ten provinces of Canada. This marks the first time that a national smoking survey of young people has been conducted in Canada.

This survey focuses on youths 10 to 19 years of age. Data for persons 10 to 14 years of age were collected through a sample of students in classrooms across Canada. Youth aged 15 to 19 were interviewed as a supplement of the Labour Force Survey. Essentially the same question content was used for the two components; however, the collection methodologies of the two components differed significantly.

This manual has been produced to facilitate the use of the microdata file of the survey results. Any questions about the data set or its use should be directed to:

Statistics Canada
Joan Coulter
Special Surveys Division, Statistics Canada
5th floor, Jean Talon Building
Tunney's Pasture
Ottawa, Ontario K1A 0T6
Telephone: (613) 951-3261
Facsimile: (613) 951-0562

Health Canada
Margaret Morin
Office of Tobacco Control
Health Canada
5th Fl. Tower "A"
11 Holland Ave
Postal Locator 3005B
Ottawa, Ontario K1A 0K9
Telephone: (613) 954-0152
Facsimile: (613) 941-1551

IT IS IMPORTANT FOR USERS TO BECOME FAMILIAR WITH THE CONTENTS OF THIS DOCUMENT BEFORE TABULATING, ANALYSING, PUBLISHING OR OTHERWISE RELEASING ANY ESTIMATES DERIVED FROM THE YOUTH SMOKING SURVEY MICRODATA FILE.

2 BACKGROUND

Since 1964, Health Canada has collected information relating to the smoking behaviour of Canadians through supplements to the Labour Force Survey (LFS) and, most recently, through the Survey on Smoking in Canada. However, most surveys on the smoking behaviour of Canadians have been limited to adults 15 years of age and older. Consequently, there existed an information gap concerning the smoking behaviour of Canadian youth. This survey addresses the need for information on youth patterns of use and attitudes towards tobacco products.

YSS and Health Canada Policies and Programs

As part of the Canadian government's recent National Action Plan to Combat Tobacco Smuggling, tobacco taxes were lowered in February 1994. Since lower prices may prompt young people to smoke more, the following legislative and regulatory changes have been or will be implemented in the near future:

- banning the manufacture of "kiddie packs" which are targeted at young buyers;
- raising the legal age for purchasing cigarettes;
- increased fines for the sale of cigarettes to minors;
- restricting locations of vending machines to bars, taverns, and other similar beverage rooms; and
- making health warnings on tobacco packaging more effective.

A comprehensive public education campaign on tobacco was also launched. Elements of the campaign included:

- a national media campaign to make young people aware of the harmful effects of smoking;
- new efforts to reach families, new parents and others who serve as role models for children;
- support for school education programs; and
- increased efforts to reach young women who are starting to smoke.

The Youth Smoking Survey (YSS) is a key component to Health Canada's building of baseline information to be used in planning and evaluating smoking behaviour.

3 OBJECTIVES

The Youth Smoking Survey (YSS) is a national survey designed to provide both national (excluding Yukon and Northwest Territories) and provincial baseline data on the knowledge, attitudes and behaviours of young people in Canada on a wide variety of tobacco issues. To support national and provincial efforts in preventing and reducing tobacco use by Canadian youth, it is necessary to obtain survey data that provide estimates of smoking prevalence in each region and enable analysis of the social factors associated with tobacco use.

The YSS was specifically designed to collect information on the following topic areas:

- the prevalence of smoking among 10-19 year olds;
- the types of smoking behaviour among young people (e.g. experimental smoking, occasional smoking, regular smoking);
- the social and demographic factors associated with smoking behaviour (e.g. what motivates young people to smoke, the influence of family and peers);
- where and how young people obtain cigarettes;
- attitudes and beliefs about smoking;
- awareness of health risks due to smoking;
- the impact of smoking policies in schools and the workplace;
- the perception of sponsorship of cultural programs and sporting events by tobacco corporations;
- brand recognition among young people and their reaction to plain packaging;
- basic household demographics including household composition and labour force activity, industry and occupation of parents.

Data from the Youth Smoking Survey will aid Health Canada to assess and develop public education programs which are intended to inform Canadian youth about the health risks associated with smoking.

This survey will also provide policy makers and researchers in the area of health promotion and health protection with accurate, current and clear summary findings.

4 CONCEPTS AND DEFINITIONS

This chapter outlines concepts and definitions of interest to the users of the Youth Smoking Survey data. The concepts and definitions used in the Labour Force Survey are described in section 4.1 while those specific to the Youth Smoking Survey are given in section 4.2. Users are referred to Chapter 12 of this document for a copy of the actual questionnaires used in the survey.

4.1 Labour Force Survey Concepts and Definitions

Labour Force Status

Status of the respondent in the labour market : a member of the non-institutional population 15 years and over is designated as either **employed**, **unemployed** or **not in the labour force**.

Employed

Employed persons are those who, during the reference week:

(a) did any work[1] at all

(b) had a job but were not at work due to:
- own illness or disability
- personal or family responsibilities
- bad weather
- labour dispute
- vacation
- other reason not specified above (excluding persons on layoff and persons whose job attachment was to a job starting at a definite date in the future).

Unemployed

Unemployed persons are those who, during the reference week:

---

[1] Work includes any work for pay or profit, that is, paid work in the context of an employer-employee relationship, or self-employment. It also includes unpaid family work where unpaid family work is defined as unpaid work which contributed directly to the operation of a farm, business or professional practice owned or operated by a related member of the household. Such activities may include keeping books, selling products, waiting on tables, and so on. Tasks such as housework or maintenance of the home are not considered unpaid family work.

(a) were without work, had actively looked for work in the past four weeks (ending with reference week), and were available for work[2];

(b) had not actively looked for work in the past four weeks but had been on layoff [3] and were available for work;

(c) had not actively looked for work in the past four weeks but had a new job to start in four weeks or less from the reference week and were available for work.

Not in the Labour Force

Those persons in the civilian non-institutional population 15 years of age and over who, during the reference week, were neither employed nor unemployed.

Out-of-scope

Those persons in the non-civilian or non-permanent residence population during the reference week.

Industry and Occupation

The Labour Force Survey provides information about the occupation and industry attachment of employed and unemployed persons, and of persons not in the labour force who have held a job in the past five years. Since 1984, these statistics have been based on the 1980 Standard Occupational Classification and the 1980 Standard Industrial Classification. Prior to 1984, the 1971 Standard Occupational Classification and the 1970 Standard Industrial Classification were used.

---

[2] Persons in this group meeting the following criteria are regarded as available:
(i) were full-time students seeking part-time work who also met condition (ii) below. (Full-time students looking for full-time work are classified as not available for work in the reference week.)
(ii) reported that there was no reason why they could not take a job in reference week, or if they could not take a job it was because of "own illness or disability", "personal or family responsibilities", or "already had a job".

[3] Persons are classified as being on layoff only when they expect to return to the job from which they were laid off.

Reference week

Entire calender week covered by the Labour Force Survey each month. It is usually the week containing the 15th day of the month. The interviews are conducted during the following week, called the Survey Week, and the labour force status determined is that of the reference week.

4.2 Youth Smoking Survey Concepts and Definitions

Definitions of smoking categories for the YSS have been based on categories proposed in the paper "Summary Report of the Workshop on Data for Monitoring Tobacco Use" by Mills, Stephens, and Wilkins, which appeared in Chronic Diseases in Canada in the summer of 1994. Some minor modifications were made to the categories proposed in the paper to adapt them to the population of 10 to 19 year-olds.

Currently smokes daily

Has smoked at least 100 cigarettes in one's lifetime, and has smoked at least one cigarette per day for each of the 30 days preceding the survey.

Currently smokes non-daily

Has smoked at least 100 cigarettes in one's lifetime, and has smoked at least one cigarette during the past 30 days, but has not smoked every day.

Currently smokes - frequency unknown

Has smoked at least 100 cigarettes in one's lifetime, and has smoked at least one cigarette during the last 30 days, but frequency unknown.

Formerly smoked daily

Has smoked 100 or more cigarettes in one's lifetime but has not smoked at all during the past 30 days, and has at some time smoked every day for 7 days in a row.

Formerly smoked non-daily

Has smoked 100 or more cigarettes in one's lifetime but has not smoked at all during the past 30 days, and has never smoked every day for 7 days in a row.

Formerly smoked - frequency unknown

Has smoked 100 or more cigarettes in one's lifetime but has not smoked at all during the past 30 days, and frequency of former smoking unknown.

Beginning to smoke

Has smoked between 1 and 99 cigarettes in one's lifetime, and has smoked in the past 30 days, but has not yet smoked 100 cigarettes.

Past experimenter

Has smoked between 1 and 99 cigarettes in one's lifetime, but has not smoked in the past 30 days.

Lifetime abstainer

Has smoked less than one whole cigarette in one's lifetime.

5 SURVEY METHODOLOGY

The Youth Smoking Survey was conducted in two parts. Data for persons 10 to 14 years of age were collected through a sample of students in classrooms across Canada, while youth aged 15 to 19 years were interviewed by telephone as a supplement to the Labour Force Survey.

5.1 LFS Supplement

The Youth Smoking Survey was administered in September 1994 to an augmented sample of dwellings in the Labour Force Survey (LFS) sample, and therefore its sample design is closely tied to that of the LFS. The LFS design is briefly described in Sections 5.1.1 to 5.1.4[4]. Sections 5.1.5 and 5.1.6 describe how the LFS Supplement Component of the Youth Smoking Survey departed from the basic LFS design.

5.1.1 Population Coverage

The target population for the YSS-LFS Supplement was all youth aged 15 to 19. Specifically excluded from this survey's coverage are full-time members of the Canadian Armed Forces, inmates of institutions, as well as residents of the Yukon and Northwest Territories, and persons living on Indian Reserves. These groups together represent an exclusion of approximately 2% of the population aged 15 and over. These exclusions were a result of using the LFS as the sampling frame for the YSS. The LFS is a monthly household survey whose sample of individuals is representative of the civilian, non-institutionalized population 15 years of age or older in Canada's ten provinces.

5.1.2 LFS Sample Design

The LFS sample is based upon a stratified, multi-stage design employing probability sampling at all stages of the design. The design principles are the same for each province. A diagram summarizing the design stages appears at the end of this section.

Primary Stratification

Provinces are first stratified into economic regions - geographic areas of a more or less homogeneous economic structure formed on the basis of federal provincial agreements. Economic regions are relatively stable over time.

---

[4] A detailed description of the LFS design is available in the Statistics Canada publication entitled **Methodology of the Canadian Labour Force Survey,** 1984-1990 (catalogue #71-526).

These economic regions are treated as primary strata and further stratification is carried out within them (see section 5.2.3).

Types of Areas

Economic regions are further disaggregated into 3 categories: self-representing areas (SRUs), non-self-representing areas (NSRUs) and special areas. Generally, SRUs are urban areas whose population as of the 1981 Census exceeds 15,000 persons or whose unique labour force characteristics demand their establishment as SRUs. For the most part, SRU boundaries are coincident with delineations established for the Census.

All SRUs in each economic region are included in the survey and, as the name implies, each is represented by its own sample.

NSRUs are the areas lying outside the SRUs and they consist largely of small urban centres and rural areas. Each economic region contains one NSRU which is represented by its own sample.

A small proportion (approximately 1%) of the LFS population is found in institutions (for example, live-in staff of hospitals or schools or permanent residents of hotels or motels), on military bases (civilian personnel only) or in remote areas of provinces which are not readily accessible to LFS interviewers. For administrative purposes, this portion of the population is sampled separately through the special area frame. This portion of the sample is selected on a province-wide basis, without reference to the stratification used for SRU and NSRU areas.

Secondary Stratification

SRU areas are next individually delineated into design strata, which reflect areas of similar socio-economic status as identified in the 1981 Census. The extent of the stratification (i.e. number of strata) depends upon the size of the SRU.

In economic regions in which the NSRU population constitutes a significant proportion of the economic region population, the NSRU is next delineated into separate urban and rural strata. Within each of these strata, further stratification is carried out to reflect differences on a number of labour force characteristics.

In special areas, strata are formed on a province-wide basis. The strata reflect the main types of special groups in the population which require special administrative sampling procedures. These are: military establishments, institutions and remote areas.

Cluster Delineation and Selection

Within each of the secondary strata found in SRU areas, a number of geographic contiguous groups of dwellings, or clusters, are formed based upon a combination of 1981 Census counts and field enumeration. These clusters generally are coincident with city blocks or block faces. The selection of a sample of clusters (generally 6 or 12 clusters) from each of these secondary strata represents the first stage of sampling in SRU areas. Within each of the secondary strata in NSRU areas, a number of large geographic areas are delineated in such a way that each one reflects the composition of the stratum within which it is located with respect to a number of socio-economic characteristics. Two or four of these areas, known as primary sampling units (or PSUs) are selected into the sample from each secondary stratum. Within each selected PSU, a number of smaller geographically contiguous groups of dwellings, or clusters, are then formed using well-defined physical features which are recognizable both on maps and in the field.

In special areas, census enumeration areas (geographic areas covered by individual enumerators for the Census) represent the first stage of selection. Within those selected, where necessary, geographically contiguous groups of dwellings or clusters are formed and the selection of a sample of these represents the second stage of sampling.

Dwelling Selection

In all three types of areas (SRU, NSRU and special areas) selected clusters are first visited by enumerators in the field and a listing of all private dwellings in the cluster is prepared. From the listing a sample of 6 dwellings (on average) is then selected. This represents the final stage of sampling.

In the 17 largest SRUs, a sample of apartments in large apartment buildings is selected from a separate register based upon information supplied by CMHC. The purpose of this is to ensure better representation of apartment dwellers in the sample as well as to minimize the effect of growth in clusters, due to construction of new apartment buildings.

Person Selection

Demographic information is obtained for all persons for whom the selected dwelling is their usual place of residence. LFS information is obtained for all civilian household members 15 years of age or older.

LFS - SAMPLE DESIGN

At every stage of the sample design, probability sampling techniques are used to ensure that the sample is random yet representative of the intended survey population.

The sample design is similar for each province.

Each province consists of a number of economic regions - areas of similar economic structure formed on the basis of federal-provincial agreements.

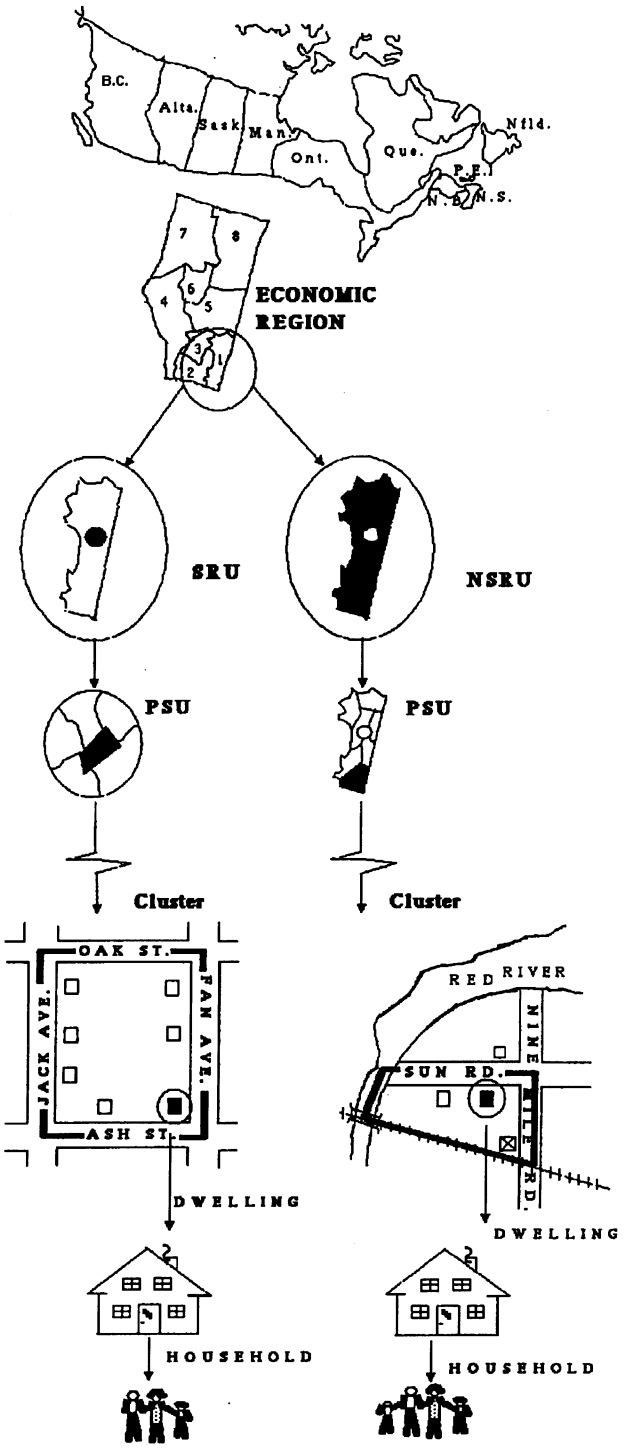
Each economic region is divided into Self-representing Units (SRU's), Non-self-representing Units (NSRU's) and Special Areas. SRU's are cities whose population exceeds 15,000 persons or whose unique characteristics demand their establishment as self-representing units. NSRU's are those areas lying outside the SRU's. Special Areas consist of military establishments, hospitals and other institutions, and remote areas.

SRU's and NSRU's are delineated into Primary Sampling Units (PSU's) which are areas that can be conveniently visited by an interviewer. A sample of PSU's is selected.

Selected PSU's are then delineated into clusters of dwellings which correspond to blocks or block faces (in urban areas) and correspond to recognizable physical boundaries (in rural areas). A sample of the clusters is selected and all private dwellings in selected clusters are listed by field enumerators.

Within each selected cluster, a sample of dwellings is selected from the list of dwellings.

Within each selected dwelling, LFS information is obtained for each civilian household member 15 years of age or older.



5.1.3 LFS Sample Size

The sample size of eligible persons in the LFS is determined so as to meet the statistical precision requirements for various labour force characteristics at the provincial and subprovincial level, to meet the requirements of federal, provincial and municipal governments as well as a host of other data users.

The monthly LFS sample consists of approximately 68,000 dwellings. After excluding dwellings found to be vacant, dwellings demolished or converted to non-residential uses, dwellings containing only ineligible persons, dwellings under construction, and seasonal dwellings, about 58,000 dwellings remain which are occupied by one or more eligible persons. From these dwellings, LFS information is obtained for approximately 112,000 civilians aged 15 or over.

5.1.4 LFS Sample Rotation

The LFS employs a panel design whereby the entire monthly sample of dwellings consists of 6 panels, or rotation groups, of approximately equal size. Each of these panels can be considered by itself to be representative of the entire LFS population. All dwellings in a rotation group remain in the LFS sample for 6 consecutive months after which time they are replaced (rotated out of the sample) by a new panel of dwellings selected from the same or similar clusters. The panels are staggered and a new panel begins each month, with each month having a panel rotating out and a new panel rotating in.

This rotation pattern was adopted to ensure that the sample of dwellings constantly reflects changes in the current housing stock and to minimize any problems of non-response or respondent burden that would occur if households were to remain in the sample for longer than 6 months. It also has the statistical advantage of providing a common sample base for short-term month-to-month comparisons of LFS characteristics.

Because of the rotation group feature, it is possible to readily conduct supplementary surveys using the LFS design by employing more or less than the full size sample.

5.1.5 Modifications to the LFS Design for the Supplement

The Youth Smoking Survey used five rotation groups from the September 1994 LFS sample, with the exception of Ontario, which used three rotation groups. In order to produce estimates at the province level, the YSS sample was increased in all provinces except Ontario and Quebec by the addition of rotation groups which had already rotated out of the LFS sample. (See the table on the following page). Although the LFS collects

information from all eligible household members aged 15 and over, the Youth Smoking Survey only collected information from those household members who were aged 15 to 19 as of September 1, 1994. Proxy responses were not permitted.

A further modification was made to the YSS sample design in order to minimize overlap with other supplementary surveys which were taking place at the same time. In one of the five rotation groups from the September 1994 LFS, youths aged 15 to 19 were interviewed only if there was no one in the household under two years of age. In three of the remaining four rotation groups from the September 1994 LFS, youths aged 15 to 19 were interviewed only if there was no one in the household under the age of 12. For the final selected rotation group from the September 1994 LFS, no restrictions applied. There were no restrictions on any of the extra rotation groups used to augment the YSS sample, which had previously rotated out of the LFS.

Province	Number of rotation groups from Sept. 94 LFS	Number of rotation groups from previous LFS	Total sampled rotation groups
Newfoundland	5	6	11
Prince Edward Island	5	6	11
Nova Scotia	5	6	11
New Brunswick	5	6	11
Quebec	5	0	5
Ontario	3	0	3
Manitoba	5	6	11
Saskatchewan	5	6	11
Alberta	5	4	9
British Columbia	5	5	10

5.1.6 Sample size by Province for the Supplement

The following table shows the number of youths aged 15 to 19 in the LFS sampled rotations who were eligible for the Youth Smoking Survey supplement.

PROVINCE	SAMPLE SIZE
Newfoundland	1,155
Prince Edward Island	447
Nova Scotia	1,186
New Brunswick	1,124
Quebec	1,444
Ontario	1,138
Manitoba	1,242
Saskatchewan	1,352
Alberta	1,288
British Columbia	1,327
CANADA	11,703

5.2 School Component

The sample design of the school component of the Youth Smoking Survey consists of a two-stage stratified clustered design in which schools are the primary sampling units and classes are the secondary units. All of the students in a selected class are included in the sample.

5.2.1 Sampling Frame

A list frame of schools was used, which covered all public and private schools in Canada. This database provided fairly complete information on the various schools' enrolment by grade level and age. The information dated from the 1991-92 school year. However, there was little geographical information concerning the schools, apart from the province in which they were located. Nor was there information on the classes in the schools.

5.2.2 Population Coverage

The target population consists of all young persons 10 to 14 years of age residing in Canada, except for those residing in the Yukon and Northwest Territories and those living in institutions or on Indian reserves. The population actually surveyed differs somewhat from the target population. The differences may be categorized as follows:

A. For operational reasons, young persons who were attending special schools (e.g. schools for the blind or for deaf-mutes) or who were attending schools located on military bases were excluded.
B. Only young persons enrolled in grades 5 to 9 were considered for the survey. This decision was based on the low number of 10-year-olds in grade 4 or of 14-year-olds in grade 10 (at the time of data collection).
C. Young persons enrolled in small classes -- classes with fewer than 10 students -- were also excluded.

It is estimated that all exclusions combined represent approximately 8% of the target population.

5.2.3 Stratification

The sample design features two levels of stratification. First, each province constitutes a main stratum. An implicit stratification by grade level (from 5 to 9 inclusive, hence 5 secondary strata per province) was used, under which the sample in each level was selected independently.

5.2.4 Sample Distribution

The school sample was distributed equally among the provinces. The requirements relating to the accuracy of the results consisted of a minimum estimable proportion (0.10) combined with a maximum coefficient of variation (16.5%) with respect to province and sex, and thus for the entire age group of 10- to 14-year-olds.

To determine firstly the sample size necessary for each province, it was necessary to make certain assumptions regarding response rates at several levels (school board, schools, parents, children), average class size and possible elimination of non-eligible students (i.e. those not meeting the age requirements). The expected response rates were based on the results of the pilot survey that had taken place the previous year. It was assumed that average class size was 30 students. For each province, it was calculated that a sample of 80 schools was sufficient to meet the accuracy requirements. The total school sample therefore consisted of 800 schools.

In each province, the sample of schools was distributed equally among the grades of interest. Accordingly, exactly 16 schools were selected for each grade.

5.2.5 Selection of Schools

The school sample was selected systematically with probability proportional to school size, that is, the total number of students for each grade. In order to guarantee good representativeness by school board size, as well as by size of school, the school file was sorted, firstly, by school board size, and secondly by size of school within each school board. Then the appropriate sampling interval was calculated and systematic selection of schools was carried out.

For Prince Edward Island, some schools were so large in terms of enrolment that their inclusion in the sample was a certainty. Technically, this amounts to creating an additional stratum for each grade.

Since the selection was done independently for each grade, a large school could be selected more than once, for different grades. This occurred especially in the small provinces. The maximum number of times that a school was thus selected was four times, in Prince Edward Island.

5.2.6 Selection of Classes

The selection of the secondary sampling units, namely classes, was done in the field. The interviewer was instructed to draw up a list of all the classes in the desired grade and carry out a selection of one of these classes using a random selection grid. All of the students in the selected class were included in the final sample. For further information on collection procedures, please see Section 6.2.

5.2.7 Treatment of Special Cases

Special cases are defined here as being all cases in which there was a change involving a selected school and affecting data collection. This includes all cases of school relocation, closure or amalgamation. In all such cases, it was important to act in such a way as to respect the probabilistic nature of the sample.

The general rule was as follows. The interviewer had to first try to determine to which school the students in the grade of interest had been relocated. He or she then checked for the following two conditions: (a) the students had been moved to a new school (one not appearing on the school list frame); and (b) no student in any other school had been

relocated to this new school. If both conditions were satisfied, then the new school replaced the old one in the sample. If not, there was no replacement.

In practice, the information concerning relocation of students was not always available or complete. The decision of whether or not to replace the school was then more complex. In all, 26 special cases were observed (representing 3.3% of the total), for which ten replacements were made.

5.2.8 Replacement for Non Response

In order to preserve the desired final sample size, we adopted a replacement strategy for schools whose agreement to participate in the sample could not be obtained. This refusal might come from either the school board or the school principal. In the first case, several schools might be affected by this decision.

In the case of a refusal by a school board, the strategy consisted of replacing that board by a school board whose enrolments for the grades of interest were the most similar. The replacement school board had to be located in the same province and obviously had to agree to participate in the survey. Next it was necessary to find schools in this replacement school board to take the place of those belonging to the original school board. Once again, the similarity of enrolments was the criterion chosen.

In the case of schools for which there was a refusal, the procedure described above was followed, whereby a replacement school from within the same school board was selected.

In total, 52 schools, or 6.5% of the sample, were replaced.

6 DATA COLLECTION

Data collection for the Youth Smoking Survey (YSS) was divided into two components: the Labour Force Survey (LFS) Component for youth aged 15 to 19 years; and the School Component for persons aged 10 to 14 years. Interviews were conducted under the voluntary provisions of the Statistics Act with the eligible youth. Proxy responses were not permitted.

6.1 LFS Supplement

Collection for the LFS supplement component was conducted during the LFS Survey Week and Post-survey Week in September 1994 (September 19 to October 1, 1994). Two weeks were necessary for data collection for two main reasons. First the sample was comprised of households from both current rotation groups and groups that had already rotated out of the LFS. This allowed interviewers to focus on collection of households in the current rotation groups during Survey Week and collection of the rotate out groups the following week. Secondly, the two-week collection period provided interviewers with another week to contact youths who were difficult to reach. (More information about the number of rotation groups selected in each province are provided in section 5.1.5)

All LFS interviewers participated in data collection. This included approximately 870 interviewers located throughout all regions of the ten provinces in Canada.

Interviewers were allowed one hour to prepare for the survey by: reading the Interviewer's Manual and completing the Review Exercise; conducting a practice interview with a family member or friend; and discussing any questions with their supervisor prior to the start of the survey.

The majority of interviews were conducted by telephone since the birth rotation group(the rotation group that is just beginning, i.e. first of six months in the LFS sample) of the LFS was excluded from the sample. The birth rotation group was excluded from the YSS sample because the first month in the LFS is conducted by personal interview. Interviews with households that did not consent to a telephone interview for the LFS were conducted through a personal visit. Only one visit was allowed for these households. Some remote areas of the country have special collection procedures for the LFS. Youth in these remote areas were only interviewed if the LFS was conducted during September. Youth in Hutterite colonies were excluded.

Interviewers were provided with an Assignment Control List that identified each household and each youth that was selected to be interviewed. Interviewers also received a labelled YSS questionnaire for each selected youth. (Note: Not all youths in the selected rotation groups were eligible to be interviewed because of overlap with a number of supplementary and longitudinal surveys. More information on this subject is provided in the section 5.1.5).

For the current rotation groups, LFS interviewers needed to deal with two collection modes: computer-assisted interviews for the LFS, and paper and pencil interviews for the School Component. Prior to LFS Survey Week, interviewers identified each household in their assignment containing one or more eligible youth by recording the survey acronym and the number of selected youth in the "Temporary Note" on their "View and Select Cases" screen.

In current rotation groups, interviewers introduced the YSS upon completion of the LFS interview. The interviewer completed a few short screening questions with the LFS household respondent to confirm that the selected youth was eligible for the YSS interview (age on September 1, 1994 was 15 to 19 years) and then asked to speak to the selected youth. Most youth were interviewed at the time of the LFS interview. If the youth was not available at the time of the LFS interview, interviewers were instructed to make no more than five telephone call-backs to contact the youth. If a youth was temporarily away from the dwelling and the household respondent could not provide a telephone number, tracing attempts were made to locate the youth. All tracing was conducted by telephone.

Collection procedures for the rotate out groups were very similar. Interviewers called the households, introduced themselves and the YSS, completed the short screening questions with a responsible member of the household, and asked to speak to the eligible youth. An LFS interview was not conducted in these households. LFS information for the household members and parents was taken from their last month of participation. Again, call-backs and tracing attempts were made to locate eligible youths.

There were three concerns relating to the telephone collection method adopted for the LFS supplement:

1. Would parents allow interviewers to speak to their children about a smoking survey?

2. Would the youth feel that the telephone interview was private, thus allowing them to answer questions freely?

3. Would it be difficult to reach youths in this age group by telephone, and would this contribute significantly to non-response?

These concerns did not prove to be problematic during data collection. The response rate for the LFS supplement was slightly higher than expected. The target was to obtain 80% response at the national level, and the actual response rate was 81.1% (see section 8.1.1 for more details on response rates to the YSS). While comments recorded on the Assignment Control List revealed that some parents did, in fact, listen to the interviews, this was rarely the case.

There was one more collection concern as to whether or not interviewers would understand and consistently apply the various eligibility rules for up to 11 selected rotation groups. However, there is no evidence of any problems in this area.

There was one collection problem which was later remedied. Youth aged 14 at the time of the LFS interview (who would have turned 15 and would be eligible for interviews) were in most cases inadvertently excluded from the selection procedures. Labels were quickly generated for these youth and collection was conducted centrally from the Regional Offices during December, 1994. No problems were encountered with the collection of these missed youth. The date for determining the age of the youth (15 - 19) and therefore the eligibility of the youth was changed to December 1 from September 1, 1994 for this group.

6.2 School Component

Collection for the School Component was conducted over a three-month period from September to November 1994. Collection was comprised of numerous activities which included mailing an introductory letter to the selected schools, administering the classroom selection procedures, administering the interviews with children in their classrooms, and completing the parent questionnaires. These collection activities were preceded by a lengthy school board approval process which began in May 1994.

Due to collection overlap with the National Population Health Survey and the National Longitudinal Survey of Children, regional offices were expected to hire new interviewers for the YSS. All regions were able to recruit experienced LFS interviewers except the Pacific region, where the majority of interviewers were new.

Recruitment was based on an average assignment size of three classrooms per interviewer. If more than one classroom, up to a maximum of four, was selected in one school, all of the classrooms were assigned to the same interviewer. This procedure ensured that each school was approached by only one interviewer and a minimal number of visits were made to the school. Recruitment was also based on the geographic distribution of the schools relative to the interviewers' residences.

Training consisted of two hours of self-study. This included reading the Interviewer's Manual and completing the Review Exercise; conducting a practice interview with a family

member or friend; and discussing any questions with their supervisor prior to the start of the survey. New interviewers in the Pacific region also received the initial training for new interviewers.

Following is a summary of the data collection process:

First Contact with School

Soon after the regional offices mailed the introductory letters to the selected schools, interviewers telephoned each school to:

- introduce the YSS to the school principal;
- obtain collaboration from the principal to participate in the survey;
- schedule an appointment for a first visit to the school; and
- verify the school address and obtain directions, if necessary.

A number of situations occurred during this initial contact that determined whether or not the class (or the school, for that matter) was within the scope of the survey. For example, what if the school no longer existed? What if the class no longer existed and the students moved to another school? What if the school principal refused to participate? A replacement strategy was developed and each case was addressed individually. (For more information refer to section 5.2.8.)

During collection, many school principals were not aware of the details of the survey because this information was not forwarded to them by their school board. One of the primary recommendations made by the school principals was that they too should have received the background package that was sent to the school boards.

First Visit to School

Upon arrival at the school, the interviewer introduced him/herself to the principal and briefly outlined the collection activities. A labelled Classroom Selection Form was used to control the collection activities. The form identified the grade that was selected for the survey. If the school had more than one class for the grade selected, the interviewer used the selection grid on the label to randomly select one of the classrooms.

The interviewer listed the name, date of birth and telephone number of each and every student in the class. For each student 10 to 14 years of age as of September 1, 1994, the interviewer prepared a package containing an introductory letter and Parental Consent Form for the student to take home. The principal or class teacher was asked to distribute and control the receipt of the completed Parental Consent Forms. The interviewer explained that he/she would return to the school in one week to pick-up the completed forms. In some

instances, the principal would not divulge the date of birth and telephone number of the students. Procedures were developed to work around this situation.

Second Visit to School

During the second visit, the interviewer:

- picked up the completed Parental Consent Forms; and
- scheduled an appointment to return to the school in one week to conduct the classroom session.

Parent Interviews

The interviewer reviewed the Parental Consent Forms to determine which children were eligible to participate in the survey. (Eligibility was dependant on the youth's age **plus** parental consent.) If a Parental Consent Form was not returned, the interviewer administered the consent procedures over the telephone. An interview was conducted by telephone with a parent or guardian of each eligible child. Most of these interviews were conducted before the classroom session.

Classroom Session (Third Visit to School)

In preparation for the classroom session, the interviewer prepared a questionnaire for each eligible child according to the preferred language of interview noted on the Parental Consent Form. The student's name was not written on the form to maintain anonymity. It was only written on an envelope for the purpose of ensuring that the correct questionnaire was given to the student.

Once in the classroom, the interviewer followed the process presented below:

- Introduced him/herself to the students.
- Explained the purpose of the survey.
- Asked the teacher to distribute the envelopes to the students.
- Distributed the Cigarette Package Recognition Hand-out face down on each desk and asked the students not to turn it over.
- Read aloud the introduction on the questionnaire.
- Explained that the hand-out should not be turned over until it was needed for the last few questions on the questionnaire.
- Completed the first 7 questions with the students to show them how to make different types of entries.
- Explained how to complete the wheel in question 19.
- Told the students not to put the completed questionnaires into the envelope but to leave them face down and separate from the envelope on their desk.
- Told students to feel free to raise their hand to ask questions.

- Asked students to complete the questionnaire.
- Answered all questions.
- Thanked students and the teacher for their co-operation and support.
- Gathered completed questionnaires, hand-outs and envelopes.

The classroom sessions, on average, lasted 30 to 40 minutes. Teachers were asked to remain in the classroom for disciplinary reasons, but were asked not to circulate among the students to protect confidentiality.

7 DATA PROCESSING

The main output of the Youth Smoking Survey is a "clean" microdata file. This section presents a brief summary of the processing steps involved in producing this file.

7.1 Data Capture

The following questionnaires were data captured in six regional offices using DC2, a UNIX based data collection and capture facility developed by Statistics Canada.

- F08 (YSS questionnaire for youth aged 15 to 19-LFS Supplement)
- F08S (YSS questionnaire for youth aged 10 to 14-School Component)
- F04S (Classroom Selection Form for School Component)
- F03S (Parent Questionnaire for School Component)

Prior to capture, completed documents were grouped into batches of 20 questionnaires. A control file was available for the LFS supplement only, as the information was extrapolated from the LFS files.

Sample verification was implemented as a quality control measure for the F08 and F08S. A sample of batches for data capture was selected and key fields were verified, i.e. re-entered and checked for discrepancies. Results of the sample verification indicated that data capture operators were well within the targeted error threshold of 2 to 3%. It was in fact estimated that the outgoing error rate at the field level was .075% for the LFS component and .260% for the School component.

Upon completion of data capture, documents were shipped to the Federal Record Centres of the National Archives.

7.2 Editing

7.2.1 Editing of the YSS-LFS Supplement (15 - 19 age group)

The editing of the YSS-LFS supplement followed a "top-down" approach. Each question was examined to verify the presence of a valid code or codes if more than one was acceptable. If none was present a "non-response" code of '9', '99', '999' or '9999' (not stated) was entered. For questions which were appropriately skipped over, a value of '6', '96', '996' or '9996' was imputed. When a skip question (i.e. a question which determined a skip pattern) was blank, the edit examined both possible flow patterns to determine the path. If determinable, the edit would impute the correct value in the skip question. If the path could not be determined, the skip question and all subsequent questions up to the next skip question were imputed with the "not stated" value of '9' or

'99'. For questions with a response item of "don't know", the value has been standardized to always appear as '7', '97' or '997'.

7.2.2 Editing of the School Component of the YSS (10 - 14 age group)

The School questionnaire was designed without any skip patterns, as it was felt that they might not be followed correctly by 10 to 14-year old respondents and might result in poor data quality. Consequently, the questions appear in a different order on the School questionnaire, compared to that of the LFS supplement, with all questions asked of all respondents. Although both questionnaires contained nearly identical question sets, the edit program used for the LFS supplement could not be repeated exactly.

The team took the decision to edit the School component using the same " top-down " logic used to edit the LFS data set. To accomplish this task, the comparable flows had to be determined before the edit program could be written. The edit program was then written to be as comparable to that of the LFS supplement as possible. The same imputed values were added where appropriate.

7.3 Coding of open-ended questions

In preparation for data capture, regional office clerks assigned codes to five items on the questionnaires. These included:

	LFS Supplement	School Component
Cigarette brand usually smoked	Q14	Q21
Cigarette brand liked best	Q21	Q24
Health problems	Q51	Q46
Health warning messages	Q54	Q48
Sponsorship of sporting and cultural events	Q59	Q52b

The source questions for these items were slightly different on the questionnaires for the LFS supplement and the School component.

A coding manual consisting of the coding procedures, sample responses, a review exercise and the code lists was developed and formed the basis for training. The coding training was conducted in a classroom setting.

The coding supervisor did a sample verification of the coded items to ensure that codes were assigned correctly.

7.3.1 Cigarette Brand

The code list for cigarette brands was provided by Health Canada. It consisted of 450 codes, 153 of which were replicated on a separate list of the top ten brands to facilitate coding of the questionnaire responses. The three-digit codes were not grouped in sequential order and varied among the different brands and descriptions.

For the LFS supplement, all responses were manually coded. The coders were required to examine the response to brand and all responses to description (i.e. regular, filter, menthol, etc.), in order to assign a code.

For the School component, only brand was obtained, not description. Therefore, if a brand listed in the questionnaire was checked, it was automatically coded and only the written responses of "other specify" needed to be coded manually.

7.3.2 Health Problems

Youth were asked to describe what health problems people can get if they smoke for many years. For the LFS supplement, interviewers recorded each answer by marking the appropriate response on the list of common health problems shown on the questionnaire, or by writing problems not covered on the list. Those responses that were specified in writing, up to a maximum of five responses, required manual coding. For the School component, the students wrote down their responses in the space provided on the questionnaire. A maximum of eight responses were manually coded.

The code list for health problems consisted of 64, two-digit codes that were grouped into major categories. Spelling errors were to be ignored, however, some responses from the school component were difficult to code due to poor spelling and penmanship.

For the microdata file, the comparable School component question was re-organized from code only variables to match the format of the LFS supplement (a list and code boxes). The codes that correspond to the list were converted to a series of check-off variables and the remaining codes were assigned to the code box variables. As the school component had three more code boxes available, even after the reorganization of the question, more than five code boxes were required. The record layout reflects (one or two) additional code boxes which would apply only to the school component as it was not possible for the LFS supplement respondents to be allowed more than five responses.

All of the responses in the "code box" variables were left justified and sorted into ascending numerical order. No duplication of codes was allowed, with the exception of code '95' for 'any other response'.

7.3.3 Health Warning Messages

Youth who indicated that they had seen health warning messages on cigarette packages, were asked to recall as many messages as they could. For the LFS supplement, interviewers would record each answer by marking the appropriate health warning message from a list. For the School component, the students were asked to write, in their own words, any health warning messages that they remembered seeing on cigarette packages. These responses were each then manually coded, to a maximum of eight responses.

Youth from both components of the survey were unlikely to provide the exact wording of the health warning messages as presented on the cigarette packages. Some key words on the questionnaire (for the LFS supplement), and on the code list (for the school component) were printed in bold letters. If a response included the words printed in bold letters, it was coded appropriately.

A few health warning messages have similar key words. For example, the word "lung" appears in the "lung cancer" and "lung disease" messages. In these cases, coders assigned the code that most closely resembled the student's response. Spelling errors were to be ignored, and as with the health problems, some responses for the School component were difficult to code due to poor spelling and penmanship.

For the microdata file, the comparable School component question was re-organized to match the format of the LFS supplement. The codes that correspond to the list of responses shown on the LFS questionnaire were converted to a series of "check off" variables, which indicate whether or not that particular health warning message was named by the respondent. The list as presented in the record layout was modified somewhat from the LFS supplement questionnaire, as the School component had one additional message in its coding structure. The last specified health warning message of "smoking can **harm children**" is applicable only to the School component.

7.3.4 Sponsorship of Sporting and Cultural Events

Youth were asked if they knew of any sporting or cultural events sponsored by tobacco corporations, and to identify all of the sponsors and events that they knew of. These responses were each manually coded, to a maximum of five responses. For the LFS supplement, interviewers were asked to record the youth's answer verbatim, and for the School component, students were to write their responses in the space provided on the questionnaire.

Lists of sponsors and events were provided by Health Canada. A three-digit coding structure was developed to code responses into two major groupings:

1. Correct match of both sponsor and event.
2. Event only given, or event with incorrect sponsor.

There were 23 codes for a match of both sponsor and event. The official name of the event was not necessary, and as the respondents were unlikely to provide the exact name of the event, key words were printed in bold letters on the code list. If a written response included the word(s) printed in bold letters, then the response was coded appropriately.

There were 18 codes for event only, or event with the incorrect sponsor. Events written by the youth that did not appear on the code list were coded to "any other event" and were considered to be incorrect responses. Spelling errors were to be ignored, but some responses were difficult to code due to poor spelling and penmanship.

7.4 Creation of Derived Variables

In order to facilitate data analysis, a number of data items (variables) on the microdata file have been derived by combining items on the questionnaire, or by an addition or other calculation. Others have been derived by assigning an 'alias' in order to avoid the disclosure of a particular cigarette brand or tobacco corporation.

The derived variables are found on the record layout immediately following the YSS questions. The derived variables are identified by variable names beginning with 'DV'.

7.5 Weighting

The principle behind estimation in a probability sample such as the LFS is that each person in the sample "represents", besides himself or herself, several other persons not in the sample. For example, in a simple random 2% sample of the population, each person in the sample represents 50 persons in the population.

The weighting phase is a step which calculates, for each record, what this number is. This weight appears on the microdata file, and must be used to derive meaningful estimates from the survey. For example, if the number of individuals who have ever smoked a whole cigarette is to be estimated, it is done by selecting the records referring to those individuals in the sample with that characteristic and summing the weights entered on those records.

Details of the method used to calculate these weights are presented in Chapter 11.

7.6 Suppression of Confidential Information

It should be noted that the 'Public Use' microdata files differ in a number of important respects from the survey 'Master' files held by Statistics Canada. These differences are the result of actions taken to protect the anonymity of individual survey respondents and tobacco corporations . Users requiring access to information excluded from the microdata files may purchase custom tabulations. Estimates generated will be released to the user, subject to meeting the guidelines for analysis and release outlined in Section 9 of this document. Data which identify specific brands or tobacco corporations may not be released to the public under any circumstances.

The Youth Smoking Survey contained two different types of information which needed to be protected to meet the confidentiality requirements of the Statistics Act:

1) information which would allow the identification of an individual respondent or the household in which he/she lived;
2) identification of specific cigarette brands or tobacco corporations.

Most variables in the respondent portion of the survey addressed characteristics which would not enable an individual respondent to be identified (for example, whether or not he/she had tried smoking, why he/she felt people his/her age started to smoke, what health problems people can get if they smoke for many years). However, some variables had to be collapsed into aggregated categories, due to low counts in the detailed categories. Examples of such variables include marital status, industry and occupation.

Information pertaining to the respondent's household also included variables for which actions were taken to prevent respondent identification. In variables for the respondent's parents, detailed categories were collapsed into aggregated categories. For example, detailed industry and occupation codes were collapsed to two-digit groupings. For other household members, information on individual persons was suppressed. Where possible, derived variables were created to summarize the characteristics of the household. For example, age of the other members is provided in summary form, indicating whether or not there are household members 0 to 9, 10 to 19, or 20+ years of age.

A number of variables identified specific cigarette brands or tobacco corporations. Measures were taken to ensure that such information was not disclosed. Variables which directly identified brands or companies were suppressed. Some other variables also had to be suppressed in order to avoid the possibility of identifying a particular brand through unique variable combinations (eg. responses on exact tar, nicotine and carbon monoxide content of cigarettes). Whenever possible, derived variables were created to summarize the information so that analysis would still be possible without disclosing confidential data. More details on the collapses and suppression of variables can be found in the record layout (See Chapter 12).

8 DATA QUALITY

8.1 Response Rates

8.1.1 LFS Component

The following table summarizes the response rates to the LFS supplement to the Youth Smoking Survey. Note that the response rate is number of individuals responding to the YSS as a percentage of number of individuals responding to LFS in rotations sampled.

	Number of respondents in Youth Smoking Survey	Person response rate to Youth Smoking Survey
Newfoundland	990	85.7%
Prince Edward Island	375	83.9%
Nova Scotia	944	79.6%
New Brunswick	866	77.1%
Québec	1,303	90.2%
Ontario	920	80.8%
Manitoba	941	75.8%
Saskatchewan	1,099	81.3%
Alberta	1,030	80.0%
British Columbia	1,023	77.1%
CANADA	9,491	81.1%

8.1.2 School Component

There were various levels of non-response for this component of the survey. A description of these levels is presented below, along with two statistical tables.

First, some degree of non-response was noted among school boards and schools. Using the substitution strategy outlined in section 5.2.8, replacements were found for all 52 refusals

(6.5% of all classes). Also described in section 5.2.7 are 26 special cases (3.3% of the total), 10 of which were replaced with other schools.

Five problem cases were encountered. For two schools in the Atlantic provinces (one in Prince Edward Island, the other in New Brunswick), parents' questionnaires could not be matched with the students' questionnaires. In another case, an error was made in grade selection at a school in Nova Scotia: a grade 7 class was chosen instead of a grade 9 class. Lastly, two Manitoba schools that had closed were replaced in the sample with two other schools even though they should not have been. In the last three cases, the schools were kept in the sample anyway, as if the selections had been made properly, and the survey weights were adjusted accordingly.

The table below shows the final distribution of classes that responded (i.e. classes where data were collected), by province and grade.

Number of responding classes by province and grade

Province	Grade					
	5	6	7	8	9	Total
Newfoundland	16	15	16	16	15	78
Prince Edward Island	16	15	16	16	16	79
Nova Scotia	16	14	16	16	15	77
New Brunswick	15	16	15	15	16	77
Quebec	16	16	16	15	16	79
Ontario	16	16	16	16	15	79
Manitoba	16	16	16	16	15	79
Saskatchewan	15	16	16	15	15	77
Alberta	16	16	16	16	16	80
British Columbia	16	16	15	16	16	79
Total	158	156	158	157	155	784

The second component of non-response relates to parents' refusal to allow their child to take part in the survey. In about 91.0% of all cases (16,262 out of a total of 17,877 eligible students), the parents agreed to let their child take the survey. An estimated 70.9% of parents

gave their consent in writing, while in the remaining cases, consent was obtained on the telephone using the procedure described in section 6.2.

Even with parental consent, there was always the possibility that the student would be absent from class at the time of the survey or would refuse to participate. In addition, a number of student questionnaires had to be rejected because they did not meet minimum quality standards. In the end, 93.6% of the eligible students for whom consent was obtained (15,217 out of 16,262) turned in an acceptable questionnaire.

There was also the possibility that the parents would refuse to complete the questionnaire or would be out each time the interviewer called. With regard to questionnaire quality, the only criterion applied was the parents' willingness to allow their responses to be shared with Health Canada. A usable questionnaire was obtained in 93.2% of all cases (15,161 out of 16,262). The number of cases in which the parents completed the questionnaire but refused to consent to data sharing was so small (under 2%) that it was decided to simply reject those questionnaires.

The last step involved matching parent and child questionnaires using a unique identifier. The total number of successfully matched questionnaires was 14,270, 93.8% of valid student questionnaires and 94.1% of valid parent questionnaires.

The overall response rates are summarized in the table below. They were calculated by dividing the final number of matched valid questionnaires by the total number of students in the sampled classes who were eligible to take part in the survey. Hence they cover all levels of non-response described above, except school non-response. The response rate for Canada as a whole was 79.8%.

Overall response rate by province

Province	Eligible students	Usable Questionnaires	Response rate (%)
Newfoundland	1815	1476	81.3
Prince Edward Island	1739	1430	82.2
Nova Scotia	1759	1431	81.4
New Brunswick	1792	1430	79.8
Quebec	2023	1556	76.9
Ontario	1785	1260	70.6
Manitoba	1622	1370	84.5
Saskatchewan	1636	1360	83.1
Alberta	1863	1516	81.4
British Columbia	1843	1441	78.2
Total	17787	14270	79.8

8.2 Survey Errors

The survey produces estimates based on information collected from and about a sample of individuals. Somewhat different estimates might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used in the survey. The difference between the estimates obtained from the sample and those resulting from a complete count taken under similar conditions is called the sampling error of the estimate.

Errors which are not related to sampling may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered on the questionnaire and errors may be introduced in the processing and tabulation of the data. These are all examples of non-sampling errors.

Over a large number of observations, randomly occurring errors will have little effect on estimates derived from the survey. However, errors occurring systematically will contribute to biases in the survey estimates. Considerable time and effort was made to reduce non-sampling errors in the survey. Quality assurance measures were implemented

at each step of the data collection and processing cycle to monitor the quality of the data. These measures included the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions, procedures to ensure that data capture errors were minimized and coding and edit quality checks to verify the processing logic.

8.2.1 Total Non Response

Total non-response can be a major source of non-sampling errors in many surveys, depending on the degree to which respondents and non-respondents differ with respect to the characteristics of interest. Total non-response occurred because the interviewer was either unable to contact the respondent, or the respondent refused to participate in the survey. Total non-response was handled by adjusting the weight of households who responded to the survey to compensate for those who did not respond.

8.2.2 Partial Non-Response

In most cases, partial non-response to the survey occurred when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information. Partial non-response is indicated by codes on the microdata file (eg. not stated).

8.2.3 Coverage

A) Coverage for LFS Component

As mentioned in Section 5.1.1 (Population Coverage), the target population for the YSS-LFS Supplement was all youth aged 15 to 19. However, specifically excluded from this survey's coverage were full-time members of the Canadian Armed Forces, inmates of institutions, as well as residents of the Yukon and Northwest Territories, and persons living on Indian Reserves. These groups together represent an exclusion of approximately 2% of the population aged 15 and over. Individuals who are members of these excluded populations may have unique characteristics that will not be reflected in the survey estimates. Users should be cautious when analyzing subgroups of the population which have characteristics that may be correlated with those population groups excluded from the YSS-LFS Supplement.

B) Coverage for School Component

As explained in section 5.2.2, the survey population and the target population are somewhat different. Students who fell into one of the excluded categories may have different tobacco-use characteristics from other students. Users studying a subgroup whose characteristics are correlated with those of excluded groups should exercise particular caution.

8.2.4 Measures of Sampling Error

Since it is an unavoidable fact that estimates from a sample survey are subject to sampling error, sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error. This section of the documentation outlines the measures of sampling error which Statistics Canada commonly uses and which it urges users producing estimates from this microdata file to use also.

The basis for measuring the potential size of sampling errors is the standard error of the estimates derived from survey results.

However, because of the large variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed relative to the estimate to which it pertains. This resulting measure, known as the coefficient of variation (C.V.) of an estimate, is obtained by dividing the standard error of the estimate by the estimate itself and is expressed as a percentage of the estimate.

For example, suppose that, based on the survey results, one estimates that 23.4% of youths aged 15 to 19 in Canada are currently cigarette smokers, and this estimate is found to have a standard error of .0043. Then the coefficient of variation of the estimate is calculated as:

$$\frac{.0043}{.234} \times 100\% = 1.8\%$$

9 GUIDELINES FOR TABULATION, ANALYSIS AND RELEASE

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

9.1 Rounding Guidelines

In order that estimates for publication or other release derived from these microdata tapes correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates:

a) Estimates in the main body of a statistical table are to be rounded to the nearest hundred units using the normal rounding technique. In normal rounding, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one. For example, in normal rounding to the nearest 100, if the last two digits are between 00 and 49, they are changed to 00 and the preceding digit (the hundreds digit) is left unchanged. If the last digits are between 50 and 99 they are changed to 00 and the preceding digit is incremented by 1.

b) Marginal sub-totals and totals in statistical tables are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units using normal rounding.

c) Averages, proportions, rates and percentages are to be computed from unrounded components (i.e. numerators and/or denominators) and then are to be rounded themselves to one decimal using normal rounding. In normal rounding to a single digit, if the final or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is increased by 1.

d) Sums and differences of aggregates (or ratios) are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units (or the nearest one decimal) using normal rounding.

e) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used resulting in estimates to be published or otherwise released which differ from corresponding estimates published by Statistics Canada, users are urged to note the reason for such differences in the publication or release document(s).

f) Under no circumstances are unrounded estimates to be published or otherwise released by users. Unrounded estimates imply greater precision than actually exists.

9.2 Sample Weighting Guidelines for Tabulation

The sample design used for the Youth Smoking Survey was not self-weighting. When producing simple estimates, including the production of ordinary statistical tables, users must apply the proper sampling weight.

If proper weights are not used, the estimates derived from the microdata tapes cannot be considered to be representative of the survey population, and will not correspond to those produced by Statistics Canada.

Users should also note that some software packages may not allow the generation of estimates that exactly match those available from Statistics Canada, because of their treatment of the weight field.

9.2.1 Definitions of types of estimates: Categorical vs. Quantitative

Before discussing how the Youth Smoking Survey data can be tabulated and analysed, it is useful to describe the two main types of point estimates of population characteristics which can be generated from the microdata file for the Youth Smoking Survey.

Categorical Estimates

Categorical estimates are estimates of the number, or percentage, of the surveyed population possessing certain characteristics or falling into some defined category. The number of youths who have ever smoked a whole cigarette or the proportion of number of days smoked in the last 30 days are examples of such estimates. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Examples of Categorical Questions :

Q: Have you ever smoked a whole cigarette?
R: Yes / No

Q: On how many of the last 30 days did you smoke one or more cigarettes?
R: None / 1 - 5 days / 6 -10 days / 11 - 20 days / 21 - 29 days / 30 days

Quantitative Estimates

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

An example of a quantitative estimate is the average number of times youth have tried to quit smoking. The numerator is an estimate of the total number of attempts to quit smoking, and its denominator is the number of youth reporting that they have attempted to quit smoking.

Examples of Quantitative Questions :

Q: How many times have you tried to quit smoking?
R: |_|_| Number of times

Q: How old were you when you first tried to quit smoking?
R: |_|_| Years

9.2.2 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained from the microdata file by summing the final weights of all records possessing the characteristic(s) of interest. Proportions and ratios of the form X/Y are obtained by:

(a) summing the final weights of records having the characteristic of interest for the numerator (X),
(b) summing the final weights of records having the characteristic of interest for the denominator (Y), then
(c) dividing the numerator estimate by the denominator estimate.

9.2.3 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata file by multiplying the value of the variable of interest by the final weight for each record, then summing this quantity over all records of interest. For example, to obtain an estimate of the total number of hours worked per week by youths who had jobs for which they were paid, multiply the value reported in Q73_63 (hours worked per week) by the final weight for the record, then sum this value over all records with Q72_62=1 (Yes, respondent has a job for which he/she gets paid).

To obtain a weighted average of the form X/Y, the numerator (X) is calculated as for a quantitative estimate and the denominator (Y) is calculated as for a categorical estimate. For example, to estimate the average number of hours worked per week by youths who had a job for which they were paid,

(a) estimate the total number of hours worked per week as described above,
(b) estimate the number of people in this category by summing the final weights of all records with Q72_62=1, then
(c) divide estimate (a) by estimate (b).

9.3 Guidelines for Statistical Analysis

The Youth Smoking Survey is based upon a complex design, with stratification and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures differs from that which is appropriate in a sample survey framework, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, analysis of variance), a method exists which can make the application of standard packages more meaningful. If the weights on the records are rescaled so that the average weight is one (1), then the results produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by dividing each weight by the overall average weight before the analysis is conducted.

In order to provide a method of assessing the quality of tabulated estimates, Statistics Canada has produced a set of Approximate Sampling Variability Tables (commonly referred to as "C.V. Tables") for the Youth Smoking Survey. These tables can be used to obtain approximate coefficients of variation for categorical-type estimates and proportions. See Chapter 10 for more details.

9.4 C.V. Release Guidelines

Before releasing and/or publishing any estimate from these microdata tapes, users should first determine the number of respondents who contribute to the calculation of the estimate. If this number is less than 30, the weighted estimate should not be released regardless of

the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the **rounded** estimate and follow the guidelines below.

Sampling Variability Guidelines

Type of Estimate	cv (in %)	Guidelines
1. Unqualified	0.0 - 16.5	Estimates can be considered for general unrestricted release. Requires no special notation.
2. Qualified	16.6 - 25.0	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning subsequent users of the high sampling variability associated with the estimates. Such estimates should be identified by the letter Q (or in some other similar fashion).
3. Confidential	25.1 - 33.3	Estimates can be considered for general unrestricted release only when sampling variabilities are obtained using an exact variance calculation procedure. Unless exact variances are obtained, such estimates should be deleted and replaced by dashes (---) in statistical tables.
4. Not for Release	33.4 or greater	Estimates cannot be released in any form under any release OR circumstances. In statistical tables, such estimates should be deleted and replaced by dashes (---).

10 APPROXIMATE SAMPLING VARIABILITY TABLES

In order to supply coefficients of variation which would be applicable to a wide variety of categorical estimates produced from this microdata file and which could be readily accessed by the user, a set of Approximate Sampling Variability Tables has been produced. These "look-up" tables allow the user to obtain an approximate coefficient of variation based on the size of the estimate calculated from the survey data.

The coefficients of variation (C.V.) are derived using the variance formula for simple random sampling and incorporating a factor which reflects the multi-stage, clustered nature of the sample design. This factor, known as the design effect, was determined by first calculating design effects for a wide range of characteristics and then choosing from among these a conservative value to be used in the look-up tables which would then apply to the entire set of characteristics.

The table below shows the design effects, sample sizes and population counts by province which were used to produce the Approximate Sampling Variability Tables.

LFS Component

PROVINCE	DESIGN EFFECT	SAMPLE SIZE	POPULATION
Newfoundland	2.02	990	48,385
Prince Edward Island	1.56	375	10,045
Nova Scotia	2.16	944	63,819
New Brunswick	1.41	866	54,934
Quebec	2.59	1303	487,428
Ontario	2.95	920	703,632
Manitoba	2.50	941	74,768
Saskatchewan	1.75	1099	71,267
Alberta	2.15	1030	185,245
British Columbia	2.45	1023	232,411
Atlantic Provinces	1.95	3175	177,183
Prairies	2.70	3070	331,280
Canada	5.47	9491	1,931,934

School Component

PROVINCE	DESIGN EFFECT	SAMPLE SIZE	POPULATION
Newfoundland	1.91	1476	44,997
Prince Edward Island	1.93	1430	9,725
Nova Scotia	1.85	1431	61,877
New Brunswick	2.40	1430	51,935
Quebec	2.66	1556	478,242
Ontario	2.05	1260	711,502
Manitoba	2.22	1370	74,755
Saskatchewan	1.94	1360	76,704
Alberta	2.16	1516	201,801
British Columbia	2.18	1441	237,721
Atlantic Provinces	2.58	5767	168,534
Prairies	2.53	4246	353,260
Canada	4.96	14,270	1,949,259

Both Components

PROVINCE	DESIGN EFFECT	SAMPLE SIZE	POPULATION
Newfoundland	2.33	2466	93,382
Prince Edward Island	2.96	1805	19,770
Nova Scotia	2.43	2375	125,696
New Brunswick	1.91	2296	106,869
Quebec	2.67	2859	965,670
Ontario	3.20	2180	1,415,134
Manitoba	2.78	2311	149,523
Saskatchewan	1.91	2459	147,971
Alberta	2.42	2546	387,046
British Columbia	2.56	2464	470,132
Atlantic Provinces	2.87	8942	345,717
Prairies	2.73	7316	684,540
Canada	5.61	23761	3,881,193

All coefficients of variation in the Approximate Sampling Variability Tables are approximate and, therefore, unofficial. Estimates of actual variance for specific variables may be obtained from Statistics Canada on a cost-recovery basis. The use of actual variance estimates would allow users to release otherwise unreleaseable estimates, i.e. estimates with coefficients of variation in the 'confidential' range.

Remember: if the number of observations on which an estimate is based is less than 30, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. This is because the formulas used for estimating the variance do not hold true for small sample sizes.

10.1 How to use the C.V. tables for Categorical Estimates

The following rules should enable the user to determine the approximate coefficients of variation from the Sampling Variability Tables for estimates of the number, proportion or percentage of the surveyed population possessing a certain characteristic and for ratios and differences between such estimates.

Rule 1: Estimates of Numbers Possessing a Characteristic (Aggregates)

The coefficient of variation depends only on the size of the estimate itself. On the Sampling Variability Table for the appropriate geographic area, locate the estimated number in the left-most column of the table (headed "Numerator of Percentage") and follow the asterisks (if any) across to the first figure encountered. This figure is the approximate coefficient of variation.

Rule 2: Estimates of Proportions or Percentages Possessing a Characteristic

The coefficient of variation of an estimated proportion or percentage depends on both the size of the proportion or percentage and the size of the total upon which the proportion or percentage is based. Estimated proportions or percentages are relatively more reliable than the corresponding estimates of the numerator of the proportion or percentage, when the proportion or percentage is based upon a sub-group of the population. For example, the proportion of "current smokers aged 15 to 19" is more reliable than the estimated number of "current smokers aged 15 to 19". (Note that in the tables the cv's decline in value reading from left to right).

When the proportion or percentage is based upon the total population of the geographic area covered by the table, the cv of the proportion or percentage is the same as the cv of the numerator of the proportion or percentage. In this case, Rule 1 can be used.

When the proportion or percentage is based upon a subset of the total population (e.g. those in a particular sex or age group), reference should be made to the proportion or percentage (across the top of the table) and to the numerator of the proportion or percentage (down the left side of the table). The intersection of the appropriate row and column gives the coefficient of variation.

Rule 3: Estimates of Differences Between Aggregates or Percentages

The standard error of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard error considered separately. That is, the standard error of a difference ($\hat{d} = \hat{X}_1 - \hat{X}_2$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The coefficient of variation of $\hat{d}$ is given by $\sqrt{\hat{d}} / \hat{d}$. This formula is accurate for the difference between separate and uncorrelated characteristics, but is only approximate otherwise.

Rule 4: Estimates of Ratios

In the case where the numerator is a subset of the denominator, the ratio should be converted to a percentage and Rule 2 applied. This would apply, for example, to the case where the denominator is the number of youth aged 15 to 19 and the numerator is the number of "current smokers aged 15 to 19".

In the case where the numerator is not a subset of the denominator, as for example, the ratio of the number of "current smokers aged 15 to 19" as compared to the number of "former smokers aged 15 to 19", the standard deviation of the ratio of the estimates is approximately equal to the square root of the sum of squares of each coefficient of variation considered separately multiplied by R. That is, the standard error of a ratio ($\hat{R} = \hat{X}_1 / \hat{X}_2$) is:

$$\sigma_{\hat{R}} = \hat{R}\sqrt{\alpha_1^2 + \alpha_2^2}$$

where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

The coefficient of variation of $\hat{R}$ is given by $\sqrt{\hat{R}} / \hat{R}$. The formula will tend to overstate the error, if $\hat{X}_1$ and $\hat{X}_2$ are positively correlated and understate the error if $\hat{X}_1$ and $\hat{X}_2$ are negatively correlated.

Rule 5: Estimates of Differences of Ratios

In this case, Rules 3 and 4 are combined. The cv's for the two ratios are first determined using Rule 4, and then the cv of their difference is found using Rule 3.

10.2 Examples of using the C.V. tables for Categorical Estimates

The following 'real life' examples are included to assist users in applying the foregoing rules.

Example 1 : Estimates of Numbers Possessing a Characteristic (Aggregates)

Suppose that a user estimates that 451,858 youths aged 15 to 19 were current smokers in the reference period. How does the user determine the coefficient of variation of this estimate?

(1) Refer to the cv table for CANADA.

(2) The estimated aggregate (451,858) does not appear in the left-hand column (the 'Numerator of Percentage' column), so it is necessary to use the figure closest to it, namely 450,000.

(3) The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry on that row, namely, 4.3%.

(4) So the approximate coefficient of variation of the estimate is 4.3%.
The finding that there were 451,858 youths aged 15 to 19 who were current smokers in the reference period is publishable with no qualifications.

Example 2 : Estimates of Proportions or Percentages Possessing a Characteristic

Suppose that the user estimates that 341,236/451,858=75.5% of youths aged 15 to 19 who were current smokers in the reference period reported that they smoke on a daily basis. How does the user determine the coefficient of variation of this estimate?

(1) Refer to the table for CANADA.

(2) Because the estimate is a percentage which is based on a subset of the total population (i.e., youths aged 15 to 19 who were current smokers), it is necessary to use both the percentage (75.5%) and the numerator portion of the percentage (341,236) in determining the coefficient of variation.

(3) The numerator, 341,236, does not appear in the left-hand column (the 'Numerator of Percentage' column) so it is necessary to use the figure closet to it, namely 350,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the figure closest to it, 70.0%.

(4) The figure at the intersection of the row and column used, namely 3.1% is the coefficient of variation to be used.

(5) So the approximate coefficient of variation of the estimate is 3.1%. The finding that 75.5% of youths aged 15 to 19 who were current smokers smoke on a daily basis can be published with no qualifications.

Example 3 : Estimates of Differences Between Aggregates or Percentages

Suppose that a user estimates that 177,052/224,852=78.7% of female youths aged 15 to 19 who were current smokers and reported that they smoke daily, while 164,184/227,006=72.3% of male youths aged 15 to 19 who were current smokers and reported that they smoke daily. How does the user determine the coefficient of variation of the difference between these two estimates?

(1) Using the CANADA cv table in the same manner as described in example 2 gives the cv of the estimate for females as 4.1%, and the cv of the estimate for males as 4.7%.

(2) Using rule 3, the standard error of a difference ($\hat{d} = \hat{X}_1 - \hat{X}_2$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

That is, the standard error of the difference $\hat{d} = (.787-.723) = .065$ is:

$$\sigma_{\hat{d}} = \sqrt{[(.787)(.041)]^2 + [(.723)(.047)]^2}$$

$$= \sqrt{(.001041) + (.001155)}$$

$$= .047$$

(3) The coefficient of variation of $\hat{d}$ is given by $\sqrt{\hat{d}} / \hat{d} = .047/.065 = 0.72$.

(4) So the approximate coefficient of variation of the difference between the estimates is 72%. This estimate can not be released under any circumstances and should be deleted and replaced by dashes.

Example 4 : Estimates of Ratios

Suppose that the user estimates that 177,052 female youths aged 15 to 19 who were current smokers and smoked daily, while 164,184 male youths aged 15 to 19 who were current smokers and smoked daily. The user is interested in comparing the estimate of women versus that of men in the form of a ratio. How does the user determine the coefficient of variation of this estimate?

(1) First of all, this estimate is a ratio estimate, where the numerator of the estimate ($= \hat{X}_1$) is the number of female youths aged 15 to 19 who were current smokers and smoked daily. The denominator of the estimate ($= \hat{X}_2$) is the number of male youths aged 15 to 19 who were current smokers and smoked daily.

(2) Refer to the table for CANADA.

(3) The numerator of this ratio estimate is 177,052. The figure closest to it is 200,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 6.9%.

(4) The denominator of this ratio estimate is 164,184. The figure closest to it is 150,000. The coefficient of variation for this estimate is found by referring to the first non-asterisk entry on that row, namely, 8.2%.

(5) So the approximate coefficient of variation of the ratio estimate is given by rule 4, which is,

$$\alpha_{\hat{R}} = \sqrt{\alpha_1^2 + \alpha_2^2}$$

where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

That is ,

$$\alpha_{\hat{R}} = \sqrt{(.069)^2 + (.082)^2}$$

$$= 0.107$$

The obtained ratio of female versus male youths aged 15 to 19 who were current smokers and smoked daily is 177,052/164,184 which is 1.08:1. The coefficient of variation of this estimate is 10.7%, which is releasable with no qualifications.

10.3 How to use the C.V. tables to obtain Confidence Limits

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

If sampling of the population is repeated indefinitely, with each sample leading to a new confidence interval for an estimate, then in 95% of the samples the interval will cover the true population value.

Using the standard error of an estimate, confidence intervals for estimates may be obtained assuming that under repeated sampling of the population, the various estimates obtained for a population characteristic are normally distributed about the true population value. Under this assumption, the chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and about 99 out 100 that the differences would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

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

Confidence intervals for an estimate can be calculated directly from the Approximate Sampling Variability Tables by first determining from the appropriate table the coefficient of variation of the estimate $\hat{X}$, and then using the following formula to convert to a confidence interval CI:

$$CI_X = [\hat{X} - t\hat{X}\alpha_{\hat{X}} , \hat{X} + t\hat{X}\alpha_{\hat{X}}]$$

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

t = 1 if a 68% confidence interval is desired
t = 1.6 if a 90% confidence interval is desired
t = 2 if a 95% confidence interval is desired
t = 3 if a 99% confidence interval is desired.

Note: Release guidelines which apply to the estimate also apply to the confidence interval. For example, if the estimate is not releasable, then the confidence interval is not releasable either.

10.4 Example of using the C.V. tables to obtain confidence limits

A 95% confidence interval for the estimated proportion of youths aged 15 to 19 who were current smokers in the reference period and reported that they smoked daily (from Example 2, section 10.2) would be calculated as follows.

$\hat{x}$ = 75.5% (or expressed as a proportion = .755)

t = 2

$\alpha_{\hat{x}}$ = 3.1% (.031 expressed as a proportion) is the coefficient of variation of this estimate as determined from the tables.

$CI_{\hat{x}_1}$ = {.755 - (2) (.755) (.031), .755 + (2) (.755) (.031)}

$CI_{\hat{x}_1}$ = {.755 - .047, .755 + .047}

$CI_{\hat{x}_1}$ = {.708, .802}

With 95% confidence it can be said that between 70.8% and 80.2% of youths aged 15 to 19 who were current smokers in the reference period smoked on a daily basis.

10.5 How to use the C.V. tables to do a t-test

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

Let $\mathbf{X_1}$ and $\mathbf{X_2}$ be sample estimates for 2 characteristics of interest. Let the standard error of the difference $\hat{X}_1 - \hat{X}_2$ be σ_d .

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

10.6 Example of using the C.V. tables to do a t-test

Let us suppose we wish to test, at 5% level of significance, the hypothesis that there is no difference between the proportion of female youths aged 15 to 19 who were current smokers and smoked daily and the proportion of male youths aged 15 to 19 who were current smokers and smoked daily. From example 3, section 10.2, the standard error of the difference between these two estimates was found to be = .047. Hence ,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{.787 - .723}{.047} = \frac{.065}{.047} = 1.38.$$

Since t = 1.38 is less than 2, it must be concluded that there is no significant difference between the two estimates at the 0.05 level of significance.

10.7 Coefficients of Variation for Quantitative Estimates

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

As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding category estimate (i.e., the estimate of the number of persons contributing to the quantitative estimate). If the corresponding category estimate is not releasable, the quantitative estimate will not be either. For example, the coefficient of variation of the "total number of weeks worked by youths aged 15 to 19 who had a job for which they were paid" would be greater than the coefficient of variation of the corresponding proportion of "youths aged 15 to 19 who had a job for which they were paid". Hence if the coefficient of variation of the proportion is not releasable, then the coefficient of variation of the corresponding quantitative estimate will also not be releasable.

Coefficients of variation of such estimates can be derived as required for a specific estimate using a technique known as pseudo replication. This involves dividing the records on the microdata files into subgroups (or replicates) and determining the variation in the estimate from replicate to replicate. Users wishing to derive coefficients of variation for

quantitative estimates may contact Statistics Canada for advice on the allocation of records to appropriate replicates and the formulae to be used in these calculations.

10.8 Release cut-off's for the Youth Smoking Survey

The minimum size of the estimate at the provincial, regional and Canada levels are specified in the table below. Estimates smaller than the minimum size given in the "Not Releasable" column may not be released under any circumstances. Note that there are three tables with release cut-offs for the Youth Smoking Survey, one for the LFS component, another for the school component, and a final table for both components. These tables are to be used with the appropriate estimates. For example, if estimates are created for the 10-14 age group only (using only the file for the school component), then the appropriate table to be consulted would be the release cut-off table for the school component. The same procedure should be followed when estimates are created for the age group 15-19 (using the file for the LFS component only), that is, consult the release table for the LFS component. However, when estimates are created for the 10-19 age group (concatenating the two files), then the third table in the series, which is the release cut-off table for both components, must be used.

Table of Release Cut-offs - LFS Component

Province	Unqualified	Qualified	Confidential	Not Releasable
Newfoundland	3,500 & +	1,500-3,500	1,000-1,500	under 1,000
Prince Edward Island	1,500 & +	1,000-1,500	500-1,000	under 500
Nova Scotia	5,000 & +	2,500-5,000	1,500-2,500	under 1,500
New Brunswick	3,000 & +	1,500-3,000	1,000-1,500	under 1,000
Quebec	33,000 & +	15,000-33,000	8,500-15,000	under 8,500
Ontario	74,000 & +	34,500-74,000	20,000-34,500	under 20,000
Manitoba	6,500 & +	3,000-6,500	1,500-3,000	under 1,500
Saskatchewan	4,000 & +	2,000-4,000	1,000-2,000	under 1,000
Alberta	13,000 & +	6,000-13,000	3,500-6,000	under 3,500
British Columbia	19,000 & +	8,500-19,000	5,000-8,500	under 5,000
Atlantic Provinces	4,000 & +	1,500-4,000	1,000-1,500	under 1,000
Prairie Provinces	10,500 & +	4,500-10,500	2,500-4,500	under 2,500
CANADA	40,000 & +	17,500-40,000	10,000-17,500	under 10,000

Table of Release Cut-offs - School Component

Province	Unqualified	Qualified	Confidential	Not Releasable
Newfoundland	2,000 & +	1,000-2,000	500-1,000	under 500
Prince Edward Island	500 & +	250-500	50-250	under 50
Nova Scotia	3,000 & +	1,500-3,000	500-1,500	under 500
New Brunswick	3,000 & +	1,500-3,000	1,000-1,500	under 1,000
Quebec	28,500 & +	12,500-28,500	7,500-12,500	under 7,500
Ontario	40,000 & +	18,000-40,000	10,500-18,000	under 10,500
Manitoba	4,000 & +	2,000-4,000	1,000-2,000	under 1,000
Saskatchewan	4,000 & +	1,500-4,000	1,000-1,500	under 1,000
Alberta	10,000 & +	4,500-10,000	2,500-4,500	under 2,500
British Columbia	12,500 & +	5,500-12,500	3,000-5,500	under 3,000
Atlantic Provinces	2,500 & +	1,000-2,500	500-1,000	under 500
Prairie Provinces	7,500 & +	3,500-7,500	2,000-3,500	under 2,000
CANADA	24,500 & +	11,000-24,500	6,000-11,000	under 6,000

Table of Release Cut-offs - Both Components

Province	Unqualified	Qualified	Confidential	Not Releasable
Newfoundland	3,000 & +	1,500-3,000	1,000-1,500	under 1,000
Prince Edward Island	1,000 & +	500-1,000	250-500	under 250
Nova Scotia	4,500 & +	2,000-4,500	1,000-2,000	under 1,000
New Brunswick	3,000 & +	1,500-3,000	1,000-1,500	under 1,000
Quebec	32,000 & +	14,000-32,000	8,000-14,000	under 8,000
Ontario	72,500 & +	32,500-72,500	18,500-32,500	under 18,500
Manitoba	6,500 & +	3,000-6,500	1,500-3,000	under 1,500
Saskatchewan	4,000 & +	2,000-4,000	1,000-2,000	under 1,000
Alberta	13,000 & +	6,000-13,000	3,500-6,000	under 3,500
British Columbia	17,500 & +	7,500-17,500	4,500-7,500	under 4,500
Atlantic Provinces	4,000 & +	2,000-4,000	1,000-2,000	under 1,000
Prairie Provinces	9,500 & +	4,000-9,500	2,500-4,000	under 2,500
CANADA	33,500 & +	14,500-33,500	8,000-14,500	under 8,000

10.9 C.V. Tables

The principles outlined in section 10.8 must also be followed when consulting the CV tables for the quality of an individual estimate. That is, when an estimate has been created for the 10-14 age group only, refer to the CV tables for the School component. When an estimate has been created for the 15-19 age group only, refer to the CV tables for the LFS component. Finally, when an estimate has been created for the age group 10-19, consult the third set of CV tables which were created for both components.

11 WEIGHTING

11.1 Weighting for the LFS component

Since the Youth Smoking Survey used a sub-sample of the LFS sample, the derivation of weights for the survey records was clearly tied to the weighting procedure used for the LFS. The LFS weighting procedure is briefly described below, and a description of the YSS weighting procedure follows immediately afterwards.

11.1.1 Weighting Procedures for the LFS

In the LFS, the final weight attached to each record is the product of the following factors: the basic weight, the cluster sub-weight, the balancing factor for non-response, the rural-urban factor and the province-age-sex ratio adjustment factor. Each is described below.

Basic Weight

In a probability sample, the sample design itself determines weights which must be used to produce unbiased estimates of the population. Each record must be weighted by the inverse of the probability of selecting the person to whom the record refers. In the example of a 2% simple random sample, this probability would be .02 for each person and the records must be weighted by 1/.02=50. Because all eligible individuals in a dwelling are interviewed (directly or by proxy), this probability is essentially the same as the probability with which the dwelling is selected.

Cluster Sub-weight

The cluster delineation is such that the number of dwellings in the sample increases very slightly with moderate growth in the housing stock. Substantial growth can be tolerated in an isolated cluster before the additional sample represents a field collection problem. However, if growth takes place in more than one cluster in an interviewer assignment, the cumulative effect of all increases may create a workload problem. In clusters where substantial growth has taken place, sub-sampling is used as a means of keeping interviewer assignments manageable. The cluster sub-weight represents the inverse of this sub-sampling ratio in clusters where sub-sampling has occurred.

Non-response

Notwithstanding the strict controls of the LFS, some non-response is inevitable, despite all the attempts made by the interviewers. The LFS non-response rate is approximately 5%. For certain types of non-response (eg. household temporarily absent, refusal), data

from a previous month's interview with the household if any, is brought forward and used as the current month's data for the household.

In other cases, non-response is compensated for by proportionally increasing the weights of responding households. The weight of each responding record is increased by the ratio of the number of households that should have been interviewed, divided by the number that were actually interviewed. This adjustment is done separately for geographic areas called balancing units. It is based on the assumption that the households that have been interviewed represent the characteristics of those that should have been interviewed. To the extent that this assumption is not true, the estimates will be somewhat biased.

Rural-urban Factor

In NSRUs without sufficient rural and urban population for explicit urban and rural strata to be formed, each primary sampling unit (PSU) is composed of both urban and rural parts. Information concerning the total population in rural and urban areas is available from the 1981 Census for each PSU as well as for each economic region (ER) in which explicit urban/rural stratification is not done. Comparison by ER with the actual 1981 rural or urban census counts indicates whether the selected PSUs over- or under-represent the respective areas. The ratio of actual rural-urban counts is divided by the corresponding estimates. These two factors are computed for each relevant ER at the time of selection of the PSUs and are entered on each sample record according to the appropriate area (rural or urban) of the NSRU. Changes in these factors are incorporated at the time of PSU rotations.

LFS Sub-Weight

The product of the previously described weighting factors is called the LFS sub-weight. All members of the same sampled dwelling have the same sub-weight.

Subprovincial and Province-Age-Sex Adjustments

The sub-weight can be used to derive a valid estimate of any characteristic for which information is collected by the LFS. In particular, estimates are produced of the total number of persons 15+ in provincial economic regions and the 24 large metropolitan areas as well as of designated age-sex groups in each of the ten provinces.

Independent estimates are available monthly for various age and sex groups by province. These are population projections based on the most recent Census data, records of births and deaths, and estimates of migration. In the final step, this auxiliary information is used to transform the sub-weight into the final weight. This is done using a linear regression model. The regression is set up to ensure that the final weights it produces sum to the census projections for the auxiliary variables, namely various age-sex groups, economic regions and census metropolitan areas.

This weighting procedure ensures consistency with external Census counts, and also ensures that every member of the economic family is assigned the same weight.

11.1.2 Weighting Procedures for the Youth Smoking Survey

The principles behind the calculation of the weights for the Youth Smoking Survey are identical to those for the LFS. However, further adjustments are made to the LFS weights in order to derive a final weight for the individual records on the Youth Smoking Survey microdata file. These adjustments are described below:

(1) An adjustment to account for the additional non-response to the supplementary survey, i.e., non-response to the Youth Smoking Survey for individuals who did respond to the LFS or for whom previous month's LFS data was brought forward. This adjustment also accounts for the weights of the youths aged 15 to 19 who were excluded from participation in the survey due to restrictions placed on the households in certain rotation groups, described in Section 5.5. The weights must be adjusted to account for these youths since they were eligible in terms of age, but were not interviewed for the YSS due to response burden considerations. Technically, these youths are not non-respondents, although they will be accounted for in the non-response adjustment. Therefore, they are not included in the response rates.

(2) An adjustment to account for the use of an augmented sample, except for Ontario and Quebec which used a sub-sample, instead of the full LFS sample.

(3) A readjustment to account for independent province-age-sex projections, after the above adjustments are made. Note that because the eligible ages are only 15 to 19, adjustments in the YSS are made by single years, instead of by age groups, as is done in the LFS.

Adjustment (1) is taken into account by multiplying the LFS subweight for each responding Youth Smoking Survey record by:

$$\frac{\textit{number of rotation groups in the LFS}}{\textit{number of rotation groups selected for the YSS}}$$

to obtain an adjusted Youth Smoking Survey sub-weight (WEIGHT2). Note that the numerator in the above equation will be six in all cases, since there are always six rotation groups in the LFS. The denominator will differ by province, since different numbers of rotation groups were selected in the various provinces (see table in 5.1.5).

At this stage the weight is comprised of three components: the LFS subweight, the non-response adjustment, and the rotation group selection adjustment. A fourth component, the province-sex-age group adjustment described below was added to improve accuracy of estimates.

As mentioned previously, independent estimates are available monthly for various age and sex groups by province. These are population projections based on the most recent Census data, records of births and deaths, and estimates of migration. Using calibration estimation, the weights are calibrated, or adjusted, by the use of auxiliary information to arrive at the final weight. The calibration is set up to ensure that the final weights it produces sum to the census projections for the auxiliary variables, namely various age-sex groups for each province. Note that unlike the LFS, the YSS used single years of age instead of age groupings. This process improves the reliability of estimates that can be produced by the Youth Smoking Survey.

Adjustment (3) is performed by the calibration estimation program. The process is similar to multiplying WEIGHT2 for each Youth Smoking Survey respondent by :

$$\frac{\textit{population total for province-sex-age } i}{\textit{sum of WEIGHT2 for survey respondents in province-sex-age } i}$$

The resulting weight (FINWT) is the final weight which appears on the Youth Smoking Survey microdata file.

> Note: Users should be cautioned that the weights are applicable to the respondent and to respondent information only, and not specifically to the household members and parents of the respondent. The sample design is such that it is not appropriate to create weighted totals for variables which refer to parent and household member information unless they are in conjunction with the respondent. For example, it is quite acceptable to find the total number of

respondents aged 15-19 who smoke and who live with one parent only. However, it would not be appropriate to compute the total number of single parents who have children aged 15-19 who smoke.

11.2 Weighting for the school component

Determining the weight associated with each record requires a number of stages, each of which involves either a level of sampling or a specific adjustment.

(1) Weight of school

The first step is to calculate the initial weight of each school selected. This is equal to the inverse of the probability of selection of the school in the stratum in question. As was described in Section 5.2.5, this probability is proportional to the number of students at the school who are enroled in the grade in question.

(2) Adjustment for nonresponse at the school level

This adjustment takes account of the fact that some schools that were selected are not in the final file. Since the total number of schools selected in each stratum was always equal to 16, the weight of the school is multiplied by the following ratio:

$$\frac{16}{\textit{number of responding schools for this stratum}}$$

(3) Adjustment for the selection of a class

This adjustment in fact corresponds to the second sampling stage, when a single class of a school is selected from among all those in the grade in question. It is simply a matter of multiplying the weight obtained in the preceding stage by the total number of classes in the school at this grade level.

(4) Adjustment for student nonresponse

This adjustment is intended to represent students for whom a final record has not been obtained, for whatever reason. The weight obtained in the previous stage is multiplied by the following ratio:

$$\frac{\textit{number of eligible students in the class}}{\textit{number of students responding in that class}}$$

(5) Adjustment by province-age-sex

This final adjustment is designed to bring the survey totals into line with the projections calculated on the basis of the census data for the target population. This adjustment is made at the province-age-sex level, thus for a total of 10*5*2=100 cells. The weight obtained in the previous stage is multiplied by the following ratio:

$$\frac{\textit{projected population for province-age-sex cell to which the record belongs}}{\textit{sum of the weights of all records belonging to that cell}}$$

12 RECORD LAYOUT

The order of the questionnaire variables in the record layout are the same as the order in which they appear in the LFS supplement. The variable name contains the question number for both components. For example, the variable "Q9_17" indicates that this is question 9 in the LFS supplement and question 17 on the school component. In some cases, the 8 byte limitation of this field restricted this numbering strategy, as in "Q11B1_19", where it is not explicit that the school component question is also a part B1.

There were specific questions asked on the LFS supplement that were not asked in the school component, and vice versa. For these cases, XX was used as the question number for the non-existing question, for example, Q14B1_XX or QXX_68A.

Variables which were created for the collapsing of confidential information begin with either a "CL" or "GRP" and appear at the end of the record layout.

The comment section below all variables includes a universe statement which indicates by question number which respondents flow through each question.

On the record layout, the multiple response questions appear as several variables, each variable with response values yes, no, valid skip, not stated. The number of variables depends on the number of possible responses (one variable for each response). When a particular response was indicated, 'yes' was checked. Otherwise, 'no', meaning not indicated was checked. 'Not stated' was checked only if no responses were indicated for that entire question, and it was a question that the respondent should have answered.

All variables have a standard set of codes to indicate missing values. The values are as follows:

6 (or 96,996, etc) if the question was skipped because of a valid response to a previous question; (i.e. not applicable);

7 (or 97,997, etc) if the response to the question was "don't know";

9 (or 99,999, etc.) will be used to indicate that the answer was "not stated".

For use in tabulation software, a short description (D card), length of 40 bytes, was written for each question.

13 QUESTIONNAIRES AND CODE SHEETS

- The LFS Supplement of the YSS (Form 08)
- The School Component of the YSS (Form 08S)
- Parent Questionnaire of the School Component (Form 03S)
- Household Record Docket (Form 03) and Code Sheet
- The Labour Force Survey Questionnaire (Form 05) and Code Sheet

13.1 The LFS Supplement to the YSS (Form 08)

The LFS supplement questionnaire (Form 08) was used to collect information on the smoking behaviour of youth aged 15-19. It was interviewer administered over the telephone.